Sharp oracle inequalities in aggregation and shape restricted regression

Pierre C. Bellec

Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : ENSAE - École nationale de la statistique et de l'administration économique

Laboratoire d'accueil : CREST - Centre d'économie, statistique et sociologie, UMR 9194 CNRS

THÈSE DE DOCTORAT ÈS MATHÉMATIQUES

Spécialité : Mathématiques fondamentales

Pierre C. Bellec

Sharp oracle inequalities in aggregation and shape restricted regression

Date de soutenance : 28 juin 2016

Après avis des rapporteurs : Richard NICKL (University of Cambridge)
Bodhisattva SEN (Columbia University)

Jury de soutenance :

Alexandre TSYBAKOV	(ENSAE & CREST) Directeur de thèse
Richard Nickl	(University of Cambridge) Rapporteur
Vladimir Koltchinskii	(GeorgiaTech) Examinateur
Philippe Rigollet	(MIT) Examinateur
Arnak Dalalyan	(ENSAE & CREST) Président du jury

NNT : 2016SACLG001

Contents

Chapter 1

Introduction and overview of the results

This thesis focuses on two areas of statistics: Aggregation of estimators and shape restricted regression. The goal of this introduction is to provide a motivation for these statistical problems, to explain how these two areas are connected and to give an overview of the results derived in the next chapters.

1.1 Aggregation of estimators: Motivating applications

If several estimators are proposed for the same statistical task, is it possible to mimic the performance of the best among them? This problem is known as the model selection type aggregation problem [83]. The definition of an estimator and of the statistical task can vary. Two statistical tasks will be studied in this thesis: density estimation and regression with fixed design. We first provide some examples and motivating applications.

1.1.1 Density estimation

For the density estimation problem, the goal is to estimate an unknown density f from i.i.d. observations that are drawn from f. For simplicity, in this introduction we focus on the univariate setting where f is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. To be more precise, $f : \mathbb{R} \to [0, +\infty)$ is a measurable function with $\int_{\mathbb{R}} f(x)dx = 1$ and we observe N i.i.d. random variables $X_1, ..., X_N$ that are drawn from f. We measure the estimation error with the integrated square risk: If $\hat{f}$ is an estimator of f based on the data $X_1, ..., X_N$, the estimation error of $\hat{f}$ is given by

$$\int_{\mathbb{R}} \left(\hat{f}(x) - f(x)\right)^2 dx.$$

Let $M \geq 2$ be an integer. Assume that we are given M estimators $\hat{f}_1, ..., \hat{f}_M$ based on $X_1, ..., X_N$. The estimators $\hat{f}_1, ..., \hat{f}_M$ are referred to as the preliminary estimators and the set $\{\hat{f}_1, ..., \hat{f}_M\}$ is sometimes called the dictionary. Our goal is to mimic the performance of the best estimator in the dictionary, and to construct a new estimator

or *aggregate* $\hat{f}$ such that

$$\mathbb{E}\int_{\mathbb{R}}\left(\hat{f}(x)-f(x)\right)^2 dx \leq C \min_{j=1,\dots,M} \mathbb{E}\int_{\mathbb{R}}\left(\hat{f}_j(x)-f(x)\right)^2 dx + \delta_{N,M}, \tag{1.1}$$

where $C \geq 1$ is a constant and $\delta_{N,M} \geq 0$ is a small quantity. The inequality (1.1) formalizes that the expected squared error of $\hat{f}$ should be smaller, up to constants, than the minimal expected squared error of the estimators in the dictionary.

Sample splitting in density estimation

We now explain a general methodology to construct an aggregate $\hat{f}$ so that (1.1) holds. This methodology is based on sample splitting and reduces the problem of aggregation of estimators to that of aggregation of deterministic functions. Let $n \geq 1$ be an integer such that $n < N$. A simple sample splitting scheme is as follows.

1. Split the data into two samples, $\{X_1, ..., X_n\}$ and $\{X_{n+1}, .., X_N\}$.
2. Using the sample $\{X_{n+1}, ..., X_N\}$, construct preliminary estimators $\{\hat{f}_1, ..., \hat{f}_M\}$.
3. Aggregate these preliminary estimators using the untouched sample $\{X_1, ..., X_n\}$.

Conditionally on $X_{n+1}, ..., X_N$, the preliminary estimators $\{\hat{f}_1, ..., \hat{f}_M\}$ can be considered as frozen. Hence, for the third step above – the aggregation step – one may use procedures that aggregate deterministic functions.

Aggregation of deterministic functions. Some of the literature on aggregation problems and Chapter 2 of the present thesis consider the problem of aggregation of deterministic functions instead of the problem of aggregation of estimators. We now explain how a result on aggregation of functions yield a result on aggregation of estimators.

Let $f_1, ..., f_M : \mathbb{R} \to \mathbb{R}$ be deterministic functions. For all $\boldsymbol{\theta} = (\theta_1, ..., \theta_M)^T \in \mathbb{R}^M$, let $f_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j f_j$. Consider the estimator $f_{\hat{\boldsymbol{\theta}}}$ where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, ..., X_n, f_1, ..., f_M)$ is an estimator valued in $\mathbb{R}^M$ and based on the observations $X_1, ..., X_n$ and the functions $f_1, ..., f_M$. It is proved in [89] that if

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, ..., X_n, f_1, ..., f_M) \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^M} \left(\int_{\mathbb{R}} f_{\boldsymbol{\theta}}(x)^2 dx - \frac{2}{n}\sum_{i=1}^n f_{\boldsymbol{\theta}}(X_i) \right), \tag{1.2}$$

then the following oracle inequality holds

$$\mathbb{E}_n \int_{\mathbb{R}} \left(f_{\hat{\boldsymbol{\theta}}}(x) - f(x)\right)^2 dx \leq \min_{j=1,\dots,M} \int_{\mathbb{R}} \left(f_j(x) - f(x)\right)^2 dx + \frac{M|f|_\infty}{n}, \tag{1.3}$$

where the expectation is taken with respect to $X_1, ..., X_n$ and $\hat{\boldsymbol{\theta}}$ is defined in (1.2). We show in Chapter 2 that $f_{\hat{\boldsymbol{\theta}}}$ is an empirical risk minimizer over the linear span of $\{f_1, ..., f_M\}$. Another example of an aggregate of deterministic functions is given by $f_{\hat{\boldsymbol{\theta}}}$ where

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, ..., X_n, f_1, ..., f_M) \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_M\}} \left(\int_{\mathbb{R}} f_{\boldsymbol{\theta}}(x)^2 dx - \frac{2}{n}\sum_{i=1}^n f_{\boldsymbol{\theta}}(X_i) \right), \tag{1.4}$$

where $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_M\}$ is the canonical basis in $\mathbb{R}^M$. In Chapter 2, it is proved that this aggregate (1.4) satisfies

$$\mathbb{E}_n \int_{\mathbb{R}} \left(f_{\hat{\boldsymbol{\theta}}}(x) - f(x)\right)^2 dx \leq \min_{j=1,...,M} \int_{\mathbb{R}} \left(f_j(x) - f(x)\right)^2 dx + c \max_{j=1,...,M} |f_j|_\infty \sqrt{\frac{\log(M)}{n}}$$

for some numerical constant $c > 0$. In Chapter 2, we study another aggregate defined by

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, ..., X_n, f_1, ..., f_M) \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Lambda^M} \left(\int_{\mathbb{R}} f_{\boldsymbol{\theta}}(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}(X_i) + \frac{1}{2} \mathrm{pen}(\boldsymbol{\theta}) \right), \tag{1.5}$$

where $\mathrm{pen}(\cdot)$ is a well-chosen penalty function and Λ^M is the simplex in $\mathbb{R}^M$. It is proved in Chapter 2 that the estimator (1.5) satisfies

$$\mathbb{E}_n \int_{\mathbb{R}} \left(f_{\hat{\boldsymbol{\theta}}}(x) - f(x)\right)^2 dx \leq \min_{j=1,...,M} \int_{\mathbb{R}} \left(f_j(x) - f(x)\right)^2 dx + c \left(|f|_\infty + \max_{j=1,...,M} |f_j|_\infty \right) \frac{\log(M)}{n}$$

for some numerical constant $c > 0$.

The aggregates $f_{\hat{\boldsymbol{\theta}}}$ with the choices $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, ..., X_n, f_1, ..., f_M)$ given in (1.2), (1.4) or (1.5) are three different estimators that aggregate the deterministic functions $f_1, ..., f_M$ using the observations $X_1, ..., X_n$. These aggregates satisfy an oracle inequality of the form

$$\mathbb{E}_n \int_{\mathbb{R}} \left(f_{\hat{\boldsymbol{\theta}}}(x) - f(x)\right)^2 dx \leq \min_{j=1,...,M} \int_{\mathbb{R}} \left(f_j(x) - f(x)\right)^2 dx + \delta_{n,M}, \tag{1.6}$$

where $\mathbb{E}_n$ denotes the expectation with respect to the observations $X_1, ..., X_n$ and $\delta_{n,M} > 0$ is a deterministic quantity.

In general, it is impossible to construct an estimator $\hat{f}$ that satisfies (1.6) with $\delta_{n,M} = 0$ over large classes of unknown densities. Such impossibility results can be proved using information theoretic lower bounds such as Le Cam's inequality, Fano's Lemma, Assouad's Lemma and their variants, cf. [99, Chapter 2] for an overview of these methods and their application to statistical lower bounds. We prove such lower bounds in Chapter 2. Such impossibility results may be surprising at first. It means that there is an unavoidable error term to pay if one wants to mimic the performance of the best function in a dictionary of M deterministic functions. This unavoidable error term can be thought of as the *price to pay* for aggregation of the deterministic functions. In the past fifteen years, a line of research has focused on characterizing the optimal price to pay for aggregation in several settings.

From aggregation of deterministic functions to aggregation of estimators. With the notation defined above, an aggregation scheme that uses sample splitting is as follows.

1. Split the data into two samples, $\{X_1, ..., X_n\}$ and $\{X_{n+1}, .., X_N\}$.

2. Using the sample $\{X_{n+1}, ..., X_N\}$, construct preliminary estimators $\{\hat{f}_1, ..., \hat{f}_M\}$.

3. Aggregate these preliminary estimators using the untouched sample $\{X_1, ..., X_n\}$ by setting $\hat{f} = \sum_{j=1}^{M} \hat{\theta}_j \hat{f}_j$ where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X, ..., X_n, \hat{f}_1, ..., \hat{f}_M)$ is one of the procedures (1.2)-(1.4)-(1.5).

By conditioning on the first sample $\{X_{n+1}, ..., X_N\}$, the aggregation result for deterministic functions (1.6) implies that

$$\mathbb{E} \int_{\mathbb{R}} \left(\hat{f} - f(x)\right)^2 dx \leq \min_{j=1,...,M} \mathbb{E}_{N-n} \int_{\mathbb{R}} \left(\hat{f}_j(x) - f(x)\right)^2 dx + \delta_{n,M},$$

where $\mathbb{E}$ denotes the expectation with respect to the complete sample

$$\{X_1, ..., X_n, X_{n+1}, ..., X_N\},$$

and $\mathbb{E}_{N-n}$ denotes the expectation with respect to the sample $\{X_{n+1}, ..., X_N\}$ used to construct the preliminary estimators. Note that in the above display, $\mathbb{E}_{N-n}$ may be replaced by $\mathbb{E}$ since $\hat{f}_j$ is independent from $\{X_1, ..., X_n\}$. Thus, data-splitting always allows us to reduce the problem of aggregation of estimators to that of aggregation of deterministic functions. We now explore two outcomes of this data-splitting strategy.

Sobolev ellipsoids and adaptation to the smoothness

For all integer $\beta \geq 1$ and positive number $L > 0$, define the set $\mathcal{S}(\beta, L)$ as the set of all densities $f : \mathbb{R} \to [0, \infty)$ with $\int_{\mathbb{R}} f(x)dx = 1$ such that f is $(\beta - 1)$-times differentiable, its derivative $f^{(\beta-1)}$ is absolutely continuous and $\int_{\mathbb{R}} \left(f^{(\beta)}(x)\right)^2 dx \leq L^2$.

Consider i.i.d. observations $X_1, ..., X_N$ drawn from an unknown density f. For any fixed integer β, there exists a kernel estimator $\hat{f}_\beta$ based on the observations $X_1, ..., X_N$ such that

$$\text{if } f \in \mathcal{S}(\beta, L) \qquad \text{then} \qquad \int_{\mathbb{R}} (\hat{f}_\beta(x) - f(x))^2 dx \leq C(\beta, L) N^{-\frac{2\beta}{2\beta+1}}$$

for some constant $C(\beta, L)$ that is independent of N. The smoothness β of the unknown density is not known in practice and a natural question is that of adaptation to β. That is, without the knowledge of β, we would like to construct an estimator $\hat{f}$ such that

$$\int_{\mathbb{R}} (\hat{f}(x) - f(x))^2 dx \leq C'(\beta, L) N^{-\frac{2\beta}{2\beta+1}}$$

for all $\beta \geq 1$ and $L > 0$ such that $f \in \mathcal{S}(\beta, L)$ for some constant $C'(\beta, L)$.

A natural solution to this problem is to construct an aggregate $\hat{f}$ that is nearly as good as any of the estimators $\{\hat{f}_\beta, \beta = 1, ..., \beta_N^{max}\}$ for some value β_N^{max} such that $\beta_N^{max} \to +\infty$ as $N \to +\infty$.

A precise construction of such aggregate $\hat{f}$ is as follows. Let $\beta^{max} = M = \lceil \log N \rceil$. Assume that N is even and let $n = N/2$. With the observations $X_{n+1}, ..., X_N$, we construct preliminary estimators $\hat{f}_1, ..., \hat{f}_M$ such that for each $j = 1, ..., M$, the estimator $\hat{f}_j$ satisfies that

$$\text{if } f \in \mathcal{S}(j, L) \qquad \text{then} \qquad \int_{\mathbb{R}} (\hat{f}_j(x) - f(x))^2 dx \leq C(j, L) n^{-\frac{2j}{2j+1}}. \tag{1.7}$$

Then, we construct the aggregate $\hat{f} = \sum_{j=1}^{M} \hat{\theta}_j \hat{f}_j$ where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, ..., X_n, \hat{f}_1, ..., \hat{f}_M)$ is defined in (1.2). If $f \in \mathcal{S}(\beta^*, L)$ for some unknown constant β^* and N is large

enough so that $M \geq \beta^*$, by (1.3) and (1.7) we have simultaneously

$$\mathbb{E}\int_{\mathbb{R}}\left(\hat{f}(x)-f(x)\right)^2 dx \leq \min_{j=1,\dots,M}\mathbb{E}\int_{\mathbb{R}}\left(\hat{f}_j(x)-f(x)\right)^2 dx + \frac{|f|_\infty M}{n}, \tag{1.8}$$

$$\text{and as } f\in\mathcal{S}(\beta^*,L), \qquad \mathbb{E}\int_{\mathbb{R}}\left(\hat{f}_{\beta^*}(x)-f(x)\right)^2 dx \leq C(\beta^*,L)n^{-\frac{2\beta^*}{2\beta^*+1}}. \tag{1.9}$$

By combining these two inequalities, we obtain

$$\mathbb{E}\int_{\mathbb{R}}\left(\hat{f}(x)-f(x)\right)^2 dx \leq C(\beta^*,L)n^{-\frac{2\beta^*}{2\beta^*+1}} + \frac{\lceil \log N\rceil |f|_\infty}{n}.$$

As $n=N/2$ and $\log(N)/n = o(N^{-\frac{2\beta^*}{2\beta^*+1}})$, this yields

$$\mathbb{E}\int_{\mathbb{R}}\left(\hat{f}(x)-f(x)\right)^2 dx \leq C'(\beta^*,L)N^{-\frac{2\beta^*}{2\beta^*+1}}$$

where $C'(\beta^*,L)$ is a constant independent of N. This simple adaptivity result can be extended and improved in two directions. First, it is also possible to define the Sobolev class $\mathcal{S}(\beta,L)$ for any positive real number β [99, see page 25-26]. A similar scheme to achieve adaptation in this context is derived in [89, Section 6]. We have restricted the presentation to $\beta\in\{1,2,3,...\}$ for simplicity.

Second, in the procedure presented above, N is even and we have performed an even split of the data. The observations $\{X_{N/2+1},...,X_N\}$ are used to construct preliminary estimators while the observations $\{X_1,...,X_{N/2}\}$ are use to aggregate these preliminary estimators. Instead of using an even split of the data, the size of these two samples can be optimized to reduce the asymptotic constant $C'(\beta^*,L)$, cf. [89, Sections 5].

Nearly as good as the kernel estimator with optimal bandwidth

Consider the sinc kernel

$$K(u)=\frac{\sin(\pi u)}{\pi u}, \qquad u\in\mathbb{R}. \tag{1.10}$$

Given a bandwidth $h>0$, a natural estimator of the unknown density f is the kernel estimator

$$\hat{f}_{N,h}(x)=\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{X_i-x}{h}\right), \qquad x\in\mathbb{R}. \tag{1.11}$$

Kernel estimators perform well in practice if the bandwidth $h>0$ is chosen properly. That is, for a density f, there exists an unknown bandwidth h^* such that the integrated squared risk $\mathbb{E}\int_{\mathbb{R}}(\hat{f}_{N,h^*}(x)-f(x))^2dx$ is small. A legitimate goal is to mimic the performance of the best among these kernel estimators. It means that we would like to construct an estimator $\hat{f}$ such that

$$\mathbb{E}\int_{\mathbb{R}}(\hat{f}(x)-f(x))^2dx \leq C\min_{h>0}\mathbb{E}\int_{\mathbb{R}}(\hat{f}_{N,h}(x)-f(x))^2dx+\delta,$$

for some constants $C\geq 1$ and $\delta>0$. This point of view is different from the one of the previous subsection where one assumes that the true density belongs to a smoothness class – such as the Sobolev ellipsoid $S(\beta,L)$ defined above – and uses the minimax

risk over this smoothness class as a benchmark. Here, no smoothness assumption is made on the true density f and the benchmark is $\min_{h>0} \mathbb{E} \int_{\mathbb{R}} (\hat{f}_{N,h}(x) - f(x))^2 dx$, which characterizes the performance of kernel estimator with the optimal bandwidth. The following inequality (1.12) is a particular case of [89, Theorem 5.1], specialized to the sinc kernel. If $K(\cdot)$ is the sinc kernel (1.10), there exists an aggregate $\hat{f}$ that satisfies

$$\mathbb{E} \int_{\mathbb{R}} (\hat{f}(x) - f(x))^2 dx \leq \left(1 + \frac{c}{\log N}\right) \min_{h>0} \mathbb{E} \int_{\mathbb{R}} (\hat{f}_{N,h}(x) - f(x))^2 dx + \frac{c(\log N)^3}{N} \quad (1.12)$$

for all $N \geq N_0$, where $c > 0$ is a numerical constant and N_0 is an integer that depends only on $|f|_\infty$ and $\int_{\mathbb{R}} K(u)^2 du$. This shows that the aggregate $\hat{f}$ is nearly as good as the best kernel estimator of the form (1.10).

The construction of the aggregate $\hat{f}$ also relies on sample splitting and the aggregation result for deterministic functions (1.3), with an additional averaging step, cf. [89, Section 4 and 5] for more details about this construction.

From structural or smoothness assumption to oracle behavior

In this thesis, we focus on the approach based on oracle inequalities such as (1.1) or (1.12). Instead of making strong smoothness or structural assumptions on the unknown density f, a dictionary of estimators is given, and the goal is to construct a new estimator that is nearly as good as any estimator in the dictionary. Being *nearly as good* as any estimator in the dictionary is an oracle behavior. The oracle is the best estimator in the collection, and the goal is to construct an estimator whose performance is close to that of the oracle.

In the literature on aggregation problems, this problem is known as the *model selection type* aggregation problem. This problem can also be stated using benchmarks. We are given a benchmark, say $\min_{j=1,\dots,M} \mathbb{E} \int_{\mathbb{R}} \left(\hat{f}_j(x) - f(x)\right)^2 dx$, and the goal is to construct a new estimator $\hat{f}$ that is nearly as good as the given benchmark, as in (1.1). This *oracle* or *benchmark* approach allows us to study aggregation procedures under minimal assumptions on the preliminary estimators and the unknown density f. For instance, the oracle inequalities satisfied by the estimators (1.2), (1.4) or (1.5) or the results of Chapter 2 hold under the assumption that $|f|_\infty$ is bounded from above by a constant, but under no other smoothness or structural assumption on the true density f. Hence, the oracle inequalities (1.1)-(1.12) contrast with traditional statistical results where one assumes that the true density has a particular smoothness or structure.

It is also possible to consider infinite collections of estimators, such as the collection $\{\hat{f}_{N,h}, h > 0\}$ that appears on the right hand side of (1.12). Here, the benchmark is $\min_{h>0} \mathbb{E} \int_{\mathbb{R}} (\hat{f}_{N,h}(x) - f(x))^2 dx$ and the aggregate $\hat{f}$ in (1.12) is nearly as good as any kernel estimator constructed with the sinc kernel.

1.1.2 Fixed design regression

A key step in the procedures presented above is the splitting of the data into two independent samples, where one sample is used to construct preliminary estimators

and the other sample is used to aggregate these preliminary estimators. We now turn to fixed design regression. In fixed design regression, splitting the data into two independent samples is in general problematic, since the observations are not identically distributed.

Assume that we have the observations

$$Y_i = \mu_i + \xi_i, \qquad i = 1, ..., n, \tag{1.13}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathbb{R}^n$ is unknown, $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$ is a subgaussian noise vector. We observe $\mathbf{y} = (Y_1, ..., Y_n)^T$ and we want to estimate $\boldsymbol{\mu}$. The values μ_i can be interpreted as the values $f(x_i)$ of an unknown regression function $f : \mathcal{X} \to \mathbb{R}$ at deterministic points $x_1, ..., x_n \in \mathcal{X}$, where $\mathcal{X}$ is an abstract set and $x_1, ..., x_n$ are known. Then, the equivalent setting is that we observe $\mathbf{y}$ along with $(x_1, \dots, x_n)$ but the values of x_i are of no interest and can be replaced by their indices if we measure the loss in a discrete norm. Namely, for any $\boldsymbol{u} \in \mathbb{R}^n$ we consider the scaled (or empirical) norm $\|\cdot\|$ defined by

$$\|\boldsymbol{u}\|^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2.$$

We will measure the error of an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ by the squared distance $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$.

As the observations (1.13) are not i.i.d., splitting the data into two independent samples cannot be achieved as simply as in the density estimation setting. If the noise vector $\boldsymbol{\xi}$ has the n-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ and the noise level $\sigma > 0$ is known, creating two independent samples is possible using the following randomization device, known as *sample cloning* [96, Lemma 2.1]. Let $\boldsymbol{g}$ be independent of $\mathbf{y}$ such that $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, I_n)$ and consider

$$\mathbf{y}^{(1)} \coloneqq \mathbf{y} + \sigma\boldsymbol{g}, \qquad \mathbf{y}^{(2)} \coloneqq \mathbf{y} - \sigma\boldsymbol{g}.$$

An explicit calculation of the covariance matrix of $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ reveals that these two random vectors are independent. Furthermore, $\mathbf{y}^{(i)} - \boldsymbol{\mu}$ has distribution $\mathcal{N}(\mathbf{0}, 2\sigma^2 I_n)$ for $i = 1, 2$. In this case, sample splitting is indeed possible, at the cost of a factor 2 in the variance. Then, as in the previous section, the sample $\mathbf{y}^{(1)}$ can be used to construct preliminary estimators and the sample $\mathbf{y}^{(2)}$ can be used to aggregate these preliminary estimators.

However, the sample cloning device is only available for Gaussian noise with known covariance matrix. If the noise is not Gaussian, another line of research studies the possibility of constructing preliminary estimators and aggregating them with the same data $\mathbf{y}$. The lack of independence between the preliminary estimators and the data used to aggregate them makes this problem more challenging than the aggregation problem under independence or the problem of aggregation of deterministic functions. Chapter 3 solves this problem in the case of linear or affine estimators. A surprising result is that the price to pay for aggregation is of the same order as if the preliminary estimators were independent of the data used for aggregation, i.e., there is no extra cost induced by the lack of independence between the preliminary estimators and the data used for aggregation. Let $A_1, ..., A_M$ be squared matrices of dimension n. For each $j = 1, ..., M$, we associate to the matrix A_j the linear estimator $\hat{\boldsymbol{\mu}}_j = A_j\mathbf{y}$. Chapter 3 suggests an estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\mathbf{y}, A_j, ..., A_M)$ that satisfies

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \le \min_{j=1,...,M} \mathbb{E}\|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}\|^2 + \frac{c\sigma^2 \log M}{n}, \tag{1.14}$$

where $c > 0$ is a numerical constant, under the only assumption that

$$\max_{j=1,\dots,M} |||A_j|||_2 \leq 1,$$

where $|||\cdot|||_2$ denotes the operator norm. Similar guarantees are proved for non-Gaussian noise in Chapter 3. The estimator $\hat{\boldsymbol{\mu}}$ is the first component of the solution of the optimization problem

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}}) \in \operatorname*{argmin}_{(\boldsymbol{u},\boldsymbol{\theta})\in\mathbb{R}^n\times\mathbb{R}^M : \boldsymbol{u}=\sum_{j=1}^M \theta_j \boldsymbol{\mu}_j} \left(\|\boldsymbol{u}-\mathbf{y}\|^2 + \sum_{j=1}^M \theta_j \left(2\sigma^2 \mathrm{Tr}(A_j) + \frac{1}{2}\|\hat{\boldsymbol{\mu}}_j - \boldsymbol{u}\|^2 \right) \right). \tag{1.15}$$

Thus, the construction of $\hat{\boldsymbol{\mu}}$ does not rely on sample splitting.

Before giving an overview of the results derived in the following chapters, let us describe some applications of aggregation methods in the fixed design regression setting

Adaptation to the number and location of jumps

Assume that the true mean $\boldsymbol{\mu}$ is piecewise constant with k pieces, or equivalently, that $\boldsymbol{\mu}$ has $k-1$ jumps. A jump is defined as an integer $i \in \{1, ..., n-1\}$ such that $\mu_i \neq \mu_{i+1}$. The locations of the jumps are unknown. If the jumps of $\boldsymbol{\mu}$ are the integers $i_1 < i_2 < ... < i_{k-1}$, then $\boldsymbol{\mu}$ belongs to the linear subspace

$$V_{i_1,\dots,i_{k-1}} = \left\{ \boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n : u_i = u_{i+1} \text{ if } i \notin \{i_1, ..., i_{k-1}\} \right\}.$$

A good random approximation of $\boldsymbol{\mu}$ is the projection of $\mathbf{y}$ onto the linear space $V_{i_1,\dots,i_{k-1}}$. This subspace has dimension k and if $P_{i_1,\dots,i_{k-1}}$ is the orthogonal projector onto this subspace, then a standard bias-variance decomposition yields

$$\mathbb{E}\|\boldsymbol{\mu} - P_{i_1,\dots,i_{k-1}}\mathbf{y}\|^2 = \min_{\boldsymbol{v}\in V_{i_1,\dots,i_{k-1}}} \|\boldsymbol{v}-\boldsymbol{\mu}\|^2 + \frac{\sigma^2 k}{n}.$$

The first term on the right hand side vanishes if the true mean $\boldsymbol{\mu}$ has not more than $k-1$ jumps and belongs to $V_{i_1,\dots,i_{k-1}}$. As the locations of the jumps of $\boldsymbol{\mu}$ are unknown, the random variable $P_{i_1,\dots,i_{k-1}}\mathbf{y}$ is not an estimator but an oracle. It is not accessible and can only serve as a benchmark.

Using the procedure $\hat{\boldsymbol{\mu}}$ of (1.14) as a building block, we will construct in Chapter 5 an estimator that satisfies a similar oracle inequality up to logarithmic factors. The performance of this estimator $\hat{\boldsymbol{\mu}}$ matches the performance of the oracle up to logarithmic factors. In the case where the number of jumps k is known, such an estimator can be constructed as follows. We aggregate the linear estimators $P_{i_1,\dots,i_{k-1}}$ for all possible values of $\{i_1, ..., i_k\}$, that is, we construct $M := \binom{n-1}{k-1}$ linear estimators of the form $P_{i_1,\dots,i_{k-1}}$ and use the procedure $\hat{\boldsymbol{\mu}}$ of (1.14) to aggregate them. Thus, we have

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{1\leq i_1<\dots<i_{k-1}\leq n} \mathbb{E}\|\boldsymbol{\mu} - P_{i_1,\dots,i_{k-1}}\mathbf{y}\|^2 + \frac{c\sigma^2 \log M}{n}, \tag{1.16}$$

$$\text{and } \forall i_1 < ... < i_{k-1} \qquad \mathbb{E}\|\boldsymbol{\mu} - P_{i_1,\dots,i_{k-1}}\mathbf{y}\|^2 = \min_{\boldsymbol{v}\in V_{i_1,\dots,i_{k-1}}} \|\boldsymbol{v}-\boldsymbol{\mu}\|^2 + \frac{\sigma^2 k}{n}. \tag{1.17}$$

For any $\boldsymbol{u} \in \mathbb{R}^n$, let $k(\boldsymbol{u}) \in \{1, ..., n\}$ be the integer such that $k(\boldsymbol{u}) - 1$ is the number of jumps of $\boldsymbol{u}$. The integer $k(\boldsymbol{u})$ is also the smallest integer l such that $\boldsymbol{u}$ is piecewise constant with l pieces. Define the class

$$\mathcal{C}_k = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{\mu} \text{ has at most } k-1 \text{ jumps } \} = \{\boldsymbol{\mu} \in \mathbb{R}^n : k(\boldsymbol{\mu}) \leq k\},$$

where $k = 1, ..., n$ is a fixed parameter. By observing that for any function $H : \mathbb{R}^n \to \mathbb{R}$,

$$\min_{\boldsymbol{v} \in \mathcal{C}_k} H(\boldsymbol{v}) = \min_{1 \leq i_1 < ... < i_{k-1} \leq n} \min_{\boldsymbol{v} \in V_{i_1,...,i_{k-1}}} H(\boldsymbol{v}),$$

we can combine the oracle inequalities (1.16)-(1.17) and the fact that $\log M \leq k \log(en/k)$ to obtain

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{v} \in \mathcal{C}_k} \|\boldsymbol{v} - \boldsymbol{\mu}\|^2 + (c+1)\sigma^2 k \log(en/k)/n.$$

The first term on the right hand side vanishes if the true mean $\boldsymbol{\mu}$ has not more than $k-1$ jumps, and in this case $\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq (c+1)\sigma^2 \log(en/k)/n$. In Chapter 5, we show that this risk bound is minimax optimal up to logarithmic factors over the class $\mathcal{C}_k$. Furthermore, the above construction will be refined in Chapter 5 so that the knowledge of k is not needed to construct the estimator.

Combining the oracle inequality (1.16) and the risk bound (1.17) is an example of a general device used to obtain risk bounds and oracle inequalities in several contexts [92, 93, 93, 14, 96]. The aggregate $\hat{\boldsymbol{\mu}}$ in (1.14) inherits the smallest risk bound among the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$, up to the additive error term $\frac{c\sigma^2 \log M}{n}$ which can be interpreted as the price to pay for aggregation of the M linear estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$.

In Chapter 6, we study a similar problem under an additional monotonicity constraint. An outcome of Chapter 6 is that we characterize the minimax rate up to logarithmic factor over the class $\mathcal{S}_n^{\uparrow} \cup \mathcal{S}_n^{\downarrow}$, where

$$\mathcal{S}_n^{\uparrow} := \{\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n : u_i \leq u_{i+1}, \quad i = 1, ..., n-1\} \tag{1.18}$$

is the set of nondecreasing sequences and $\mathcal{S}_n^{\downarrow} = -\mathcal{S}_n^{\uparrow}$ is the set of non-increasing sequences. Let $\hat{\boldsymbol{\mu}}^{\text{LS}}(K) = \text{argmin}_{\boldsymbol{v} \in K} \|\mathbf{y} - \boldsymbol{v}\|^2$ be the Least Squares Estimator over K for any set $K \subset \mathbb{R}^n$ and define the estimator $\hat{\boldsymbol{\mu}}$ as the first component of the solution of the optimization problem

$$(\hat{\boldsymbol{\mu}}, \hat{\theta}_{\uparrow}, \hat{\theta}_{\downarrow}) \in \underset{(\boldsymbol{u}, \theta_{\uparrow}, \theta_{\downarrow}) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \,:\, \boldsymbol{u} = \theta_{\uparrow} \hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow}) + \theta_{\downarrow} \hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\downarrow})}{\text{argmin}} H(\boldsymbol{u}, \theta_{\uparrow}, \theta_{\downarrow}),$$

$$\text{where} \quad H(\boldsymbol{u}, \theta_{\uparrow}, \theta_{\downarrow}) = \|\boldsymbol{u} - \mathbf{y}\|^2 + \frac{1}{2}\left(\hat{\theta}_{\uparrow}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{u}\|^2 + \hat{\theta}_{\downarrow}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\downarrow}) - \boldsymbol{u}\|^2\right).$$

The estimator $\hat{\boldsymbol{\mu}}$ aggregates the Least Squares estimators $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow})$ and $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\downarrow})$. We will prove in Chapters 5 and 6 that for any $k_0 \in \{1, ..., n\}$ and for all $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow} \cup \mathcal{S}_n^{\downarrow}$,

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{K \in \{\mathcal{S}_n^{\uparrow}, \mathcal{S}_n^{\downarrow}\}} \mathbb{E}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}\|^2 + \frac{4\sigma^2 \log(en)}{n}, \tag{1.19}$$

$$\forall K \in \{\mathcal{S}_n^{\uparrow}, \mathcal{S}_n^{\downarrow}\}, \qquad \mathbb{E}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in K} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k(\boldsymbol{u}) \log \frac{en}{k(\boldsymbol{u})}}{n} \tag{1.20}$$

$$\exists \boldsymbol{u} \in \mathcal{S}_n^{\uparrow} \cup \mathcal{S}_n^{\downarrow} : \qquad k(\boldsymbol{u}) = k_0 \qquad \text{and} \qquad \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \leq \frac{V^2}{k_0^2}, \tag{1.21}$$

where $V = |\mu_n - \mu_1|$ is the total variation of $\boldsymbol{\mu}$. The oracle inequality (1.19) describes the aggregation property of the aggregate $\hat{\boldsymbol{\mu}}$, (1.20) bounds from above the risk of the estimators $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ and $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\downarrow})$, while (1.21) is a deterministic approximation result. Combining (1.19)-(1.20)-(1.21) with $k_0 = \lceil (V/\sigma)^{2/3} n^{1/3} \rceil$, we obtain that

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq c\sigma^2 \log(en) \left(\left(\frac{V}{\sigma n} \right)^{2/3} + \frac{1}{n} \right)$$

for some numerical constant $c > 0$. In Chapters 5 and 6, we show that this rate is minimax optimal up to logarithmic factors.

Screening in high-dimensional statistics

A similar strategy was used in the context of high-dimensional linear regression to obtain sparsity oracle inequalities [37, 34, 92, 93]. This technique is often referred to as *sparsity pattern aggregation* or *exponential screening*, and it leads to prediction guarantees that improve upon the prediction bounds satisfied by ℓ_1-regularized estimators such as the Lasso and the Dantzig selector, cf. [92, 93] or Chapter 3 below. Oracle inequalities such as (1.14) above are used as building blocks to obtain such results.

Consider a design matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$ with p columns, and let $s \in \{1, ..., p\}$ be a fixed parameter. For any $J \subseteq \{1, ..., p\}$, let $P_J \in \mathbb{R}^{n \times n}$ be the orthogonal projector onto the linear span of the columns of $\mathbb{X}$ whose indices belong to J. Let

$$\mathcal{J} = \{ J \subseteq \{1, ..., p\} : \quad |J| = s \quad \text{and} \quad \operatorname{rank} P_J = s \}, \qquad M = |\mathcal{J}|.$$

Define the collection of linear estimators $(\hat{\boldsymbol{\mu}}_j)_{j=1,...,M}$ as the collection $(P_J \mathbf{y})_{J \in \mathcal{J}}$ and let $\hat{\boldsymbol{\mu}}$ be the estimator that aggregates the linear estimators $(\hat{\boldsymbol{\mu}}_j)_{j=1,...,M}$ so that (1.14) holds. Then we have simultaneously

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{J \in \mathcal{J}} \ \mathbb{E}\|P_J \mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 + \frac{c\sigma^2 \log M}{n}, \tag{1.22}$$

$$\text{for all } J \in \mathcal{J}, \qquad \mathbb{E}\|P_J \mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 = \min_{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0 \text{ if } j \notin J} \|\mathbb{X}\boldsymbol{\beta} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 s}{n}, \tag{1.23}$$

where the second line follows from a simple orthogonal decomposition and the fact that $\operatorname{rank} P_J = s$. Combining the oracle inequalities (1.22)-(1.23) and the fact that $\log M \leq s \log(ep/s)$, we obtain

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\beta}|_0 \leq s} \|\mathbb{X}\boldsymbol{\beta} - \boldsymbol{\mu}\|^2 + \frac{(c+1)\sigma^2 s \log(ep/s)}{n}, \tag{1.24}$$

since $\min_{J \in \mathcal{J}} \min_{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0 \text{ if } j \notin J} = \min_{\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\beta}|_0 \leq s}$. Here, $|\boldsymbol{\beta}|_0$ denotes the number of nonzero coefficients of any $\boldsymbol{\beta} \in \mathbb{R}^p$. Inequality (1.24) is called a *sharp oracle inequality* or a *regret bound*, since the constant in front of $\min_{\boldsymbol{\beta} \in \mathbb{R}^p : |\boldsymbol{\beta}|_0 \leq s} \|\mathbb{X}\boldsymbol{\beta} - \boldsymbol{\mu}\|^2$ is 1. Chapter 5 defines a related quantity, the *minimax regret*, and discusses the optimality of such regret bounds in a minimax sense. For some design matrices and most values of the parameters s, σ^2, n and p, the regret bound (1.24) is optimal in a minimax sense over the class $\{\mathbb{X}\boldsymbol{\beta} : |\boldsymbol{\beta}|_0 \leq s\}$, cf. [92].

Adaptation to the smoothness

The sharp oracle inequality (1.14) provides a general strategy to prove adaptivity results with asymptotically exact minimax constant. Consider the regression setting (1.13) with

$$\mu_i = f(i/n), \qquad i = 1, ..., n,$$

where $f : [0, 1] \to \mathbb{R}$ is an unknown function. We will write $\boldsymbol{\mu}_f = (f(1/n), ..., f(n/n))^T$ to emphasize the dependence on f. For any positive integer $\beta \in \{1, 2, ...\}$, consider the Sobolev class

$$W(\beta, L) = \left\{ f : [0, 1] \to \mathbb{R} : f^{(\beta-1)} \text{ is abs. continuous and } \int_{[0,1]} f^{(\beta)}(x)^2 dx \leq L \right\}.$$

For any $\beta \in \{1, 2, ...\}$ and any $n \geq 1$, there exists a squared matrix $A_\beta \in \mathbb{R}^{n\times n}$ that depends on β such that

$$\text{if } f \in W(\beta, L) \qquad \text{then} \qquad \mathbb{E}\|\boldsymbol{\mu}_f - A_\beta \mathbf{y}\|^2 \leq (1 + \rho_n) C^*(\beta, L) n^{-\frac{2\beta}{2\beta+1}}, \quad (1.25)$$

where ρ_n is a quantity that tends to 0 as n goes to infinity, and the constant $C^*(\beta, L)$ is the asymptotically exact minimax constant. The asymptotically exact minimax constant satisfies

$$n^{\frac{2\beta}{2\beta+1}} \inf_{\hat{\boldsymbol{\mu}}} \sup_{f\in W(\beta,L)} \mathbb{E}\|\boldsymbol{\mu}_f - \hat{\boldsymbol{\mu}}\|^2 \to C^*(\beta, L) \qquad \text{as} \quad n \to +\infty,$$

where the infimum on the left hand side is taken over all estimators. Such matrices A_β are given, for instance, by the Pinsker filters and (1.25) is Pinsker's theorem [99, Theorem 3.2].

Adaptation with respect to β is obtained in a similar fashion as in Section 1.1.1. The main difference is that here, data-splitting is not necessary since the aggregation result (1.14) allows us to aggregate the collection of estimators $(A_\beta \mathbf{y})_{\beta=1,...,M}$ with the same data as that used to construct these estimators. As a result, we obtain an adaptation result with the asymptotically exact minimax constant. Indeed, let $M = n$ and consider the estimator $\hat{\boldsymbol{\mu}}$ that aggregates the linear estimators $(\hat{\boldsymbol{\mu}}_j)_{j=1,...,M} = (A_\beta \mathbf{y})_{\beta=1,...,M}$ so that (1.14) holds. Inequalities (1.25) and (1.14) can be rewritten as

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_f\|^2 \leq \min_{\beta=1,...,M} \mathbb{E}\|A_\beta \mathbf{y} - \boldsymbol{\mu}_f\|^2 + \frac{c\sigma^2 \log M}{n}, \quad (1.26)$$

$$\text{if } f \in W(\beta, L), \qquad \mathbb{E}\|A_\beta \mathbf{y} - \boldsymbol{\mu}_f\|^2 \leq (1 + \rho_n) C^*(\beta, L) n^{-\frac{2\beta}{2\beta+1}}. \quad (1.27)$$

If $f \in W(\beta^*, L)$ for some unknown $\beta^* \in \{1, 2, ...\}$, by combining (1.26)-(1.27) we obtain that for all $n \geq \beta^*$,

$$\begin{aligned}\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_f\|^2 &\leq (1 + \rho_n) C^*(\beta^*, L) n^{-\frac{2\beta^*}{2\beta^*+1}} + \frac{c\sigma^2 \log n}{n}, \\ &= (1 + \rho_n + \delta_n) C^*(\beta^*, L) n^{-\frac{2\beta^*}{2\beta^*+1}},\end{aligned}$$

where δ_n is a sequence that tends to 0 as n goes to infinity. Thus, the aggregate $\hat{\boldsymbol{\mu}}$ achieves adaptation with the asymptotically exact minimax constant.

The Sobolev class $W(\beta, L)$ can be defined for continuous values of β (cf. [99, Definition 2.12]) and if $\beta \in (0, +\infty)$, (1.25) still holds for a matrix A_β that depends

only on n and β. In Chapter 3, we extend the construction of the previous paragraph to the case $\{\beta \in [1, +\infty)\}$. Here is an outline of this construction. Let $M \geq 2$ and $1 = \beta_1 < \beta_2 < ... < \beta_M$ be quantities that only depend on n such that

$$\beta_M \to +\infty \text{ as } n \to +\infty \qquad \text{and} \qquad \forall j = 1, ..., M-1, \qquad n^{-\frac{2\beta_j}{2\beta_j+1}} \leq (1+a_n) n^{-\frac{2\beta_{j+1}}{2\beta_{j+1}+1}}, \tag{1.28}$$

where a_n is a sequence that depends only on n and that tends to 0 as n goes to infinity. Let $\hat{\boldsymbol{\mu}}$ be the aggregate (1.15) where $(\hat{\boldsymbol{\mu}}_j)_{j=1,...,M} = (A_{\beta_j}\mathbf{y})_{j=1,...,M}$ so that (1.14) holds. Then we have

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_f\|^2 \leq \min_{j=1,...,M} \quad \mathbb{E}\|A_{\beta_j}\mathbf{y} - \boldsymbol{\mu}_f\|^2 + \frac{c\sigma^2 \log M}{n}, \tag{1.29}$$

$$\text{if } f \in W(\beta_j, L), \qquad \mathbb{E}\|A_{\beta_j}\mathbf{y} - \boldsymbol{\mu}_f\|^2 \leq (1+\rho_n) C^*(\beta, L) n^{-\frac{2\beta_j}{2\beta_j+1}}, \tag{1.30}$$

$$\text{if } \beta_j \leq \beta^* < \beta_{j+1} \text{ then} \quad n^{-\frac{2\beta_j}{2\beta_j+1}} \leq (1+a_n) n^{-\frac{2\beta_{j+1}}{2\beta_{j+1}+1}} \leq (1+a_n) n^{-\frac{2\beta^*}{2\beta^*+1}}. \tag{1.31}$$

Assume that $f \in W(\beta^*, L)$ for some unknown $\beta^* \in [1, +\infty)$. By combining (1.29)-(1.30)-(1.31), we obtain that for n large enough (so that $\beta^* < \beta_M$), we have

$$\begin{aligned}\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_f\|^2 &\leq (1+\rho_n)(1+a_n) C^*(\beta, L) n^{-\frac{2\beta^*}{2\beta^*+1}} + \frac{c\sigma^2 \log n}{n}, \\ &= (1 + \rho_n + a_n + \rho_n a_n + \delta_n) C^*(\beta^*, L) n^{-\frac{2\beta^*}{2\beta^*+1}},\end{aligned}$$

where ρ_n, a_n, δ_n tend to 0 as n goes to infinity and their roles are as follows:

- δ_n represents the price paid for aggregation,
- $(1+\rho_n)$ is the ratio between the risk of the estimator $A_\beta \mathbf{y}$ and the minimax oracle when $f \in W(\beta, L)$,
- a_n controls that the grid $\{\beta_1, ..., \beta_M\}$ is thin enough.

Thus, if we can find a grid $\beta_1 < ... < \beta_M$ such that (1.28) holds, the estimator $\hat{\boldsymbol{\mu}}$ achieves adaptation over $\{\beta \in [1, +\infty)\}$ with the asymptotically exact minimax constant. An example of a grid satisfying (1.28) is given in Chapter 3.

Dalalyan and Salmon [35] proposed a different approach to achieve a similar goal using the Exponential Weights aggregate.

1.1.3 Oracle inequalities as building blocks

We have presented several applications of the aggregation results derived in the present thesis. In the previous subsections, we emphasized that oracle inequalities may be thought of as "building blocks" in order to obtain statistical results in several settings.

- In density estimation, the oracle inequality (1.8) and the risk bound (1.9) were combined to obtain an adaptivity result with respect to Sobolev classes indexed by integer parameters.
- The oracle inequalities (1.16) and (1.17) were combined to achieve adaptation with respect to the location of jumps in isotonic regression.

- The oracle inequalities (1.19), the risk bound (1.20) and the approximation (1.21) were combined to achieve the nonparametric rate $n^{-2/3}$ is isotonic regression where the direction of monotonicity is unknown.
- The oracle inequalities (1.22) and (1.23) were combined to obtain the sparsity oracle inequality (1.24) in high-dimensional linear regression.
- The oracle inequalities (1.26) and (1.27) were combined to achieve adaptation with respect to a discrete family of Sobolev ellipsoids.
- Finally, the oracle inequality (1.29), the risk bound (1.30) and the grid approximation (1.31) were combined to achieve adaptation with respect to a continuous family of Sobolev ellipsoids.

This general idea was used to obtain many results in the past decade. An impressive outcome of this method lies in high-dimensional statistics [92, 93, 96] where the resulting sparsity oracle inequality improves substantially upon the sparsity oracle inequalities satisfied by ℓ_1-regularized methods such as the Lasso and the Dantzig selector.

1.2 Overview of the results

1.2.1 A penalized procedure over the simplex

Consider a Hilbert space $\mathcal{H}$ with a norm $\|\cdot\|_{\mathcal{H}}$. Let $h^* \in \mathcal{H}$ be some ground truth that we would like to estimate. Assume that there is some empirical criterion $R_n(\cdot)$, such that for all $h \in \mathcal{H}$,

$$\mathbb{E}[R_n(h)] = \|h - h^*\|_{\mathcal{H}}^2.$$

Then it is natural to use R_n in order to estimate h^* from the data, for instance by minimizing the data-driven criterion R_n over $\mathcal{H}$ or over a subset of $\mathcal{H}$. One may also minimize the sum of the criterion R_n plus an additive penalty in order to achieve some kind of regularization.

Let $h_1, ..., h_M$ be elements of $\mathcal{H}$. Our goal is to mimic the performance of the best element in $\{h_1, ..., h_M\}$. Ideally, we are looking for an estimator $\hat{h}$ such that

$$\mathbb{E}\|\hat{h} - h^*\|_{\mathcal{H}}^2 \leq \min_{j=1,...,M} \|h_j - h^*\|_{\mathcal{H}}^2 + \delta_{n,M}, \tag{1.32}$$

where $\delta_{n,M}$ is some small quantity. Chapters 2 and 3 study this problem. For all $\boldsymbol{\theta} = (\theta_1, ..., \theta_M)^T \in \mathbb{R}^M$, define $h_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j h_j$. A central object of these chapters is the penalty

$$\mathrm{pen}(\boldsymbol{\theta}) = \sum_{j=1}^M \theta_j \|h_j - h_{\boldsymbol{\theta}}\|_{\mathcal{H}}^2, \qquad \boldsymbol{\theta} = (\theta_1, ..., \theta_M)^T \in \Lambda^M \tag{1.33}$$

where Λ^M is the simplex in $\mathbb{R}^M$. An estimator is constructed by minimizing the penalized empirical criterion as follows. Let

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Lambda^M} \left(R_n(h_{\boldsymbol{\theta}}) + \frac{1}{2}\mathrm{pen}(\boldsymbol{\theta}) \right), \tag{1.34}$$

and define the estimator $\hat{h}$ by $\hat{h} = h_{\hat{\boldsymbol{\theta}}}$. Chapters 2 and 3 prove that this estimator satisfies (1.32) for a quantity $\delta_{n,M}$ of order $\log(M)/n$ in several settings.

Aggregation of deterministic vectors in fixed design regression. Here, let $\mathcal{H} = \mathbb{R}^n$ and consider the regression model (1.13). We observe $\mathbf{y}$ and the goal is to estimate $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y}]$. For any deterministic vector $\boldsymbol{b} \in \mathbb{R}^n$, the sum of squares criterion provides an unbiased estimate of the risk. Indeed, for all $\boldsymbol{b} \in \mathbb{R}^n$ we have

$$\mathbb{E}[\|\mathbf{y} - \boldsymbol{b}\|^2] = \|\boldsymbol{\mu} - \boldsymbol{b}\|^2 + \left(\mathbb{E}\|\mathbf{y}\|^2 - \|\boldsymbol{\mu}\|^2\right)$$

and the term $(\mathbb{E}\|\mathbf{y}\|^2 - \|\boldsymbol{\mu}\|^2)$ is independent of $\boldsymbol{b}$. Let $\boldsymbol{b}_1, ..., \boldsymbol{b}_M$ be deterministic vectors in $\mathbb{R}^n$ and let $\boldsymbol{b}_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j \boldsymbol{b}_j$ for all $\boldsymbol{\theta} = (\theta_1, ..., \theta_M)^T \in \mathbb{R}^M$. Here, the procedure (1.34) becomes

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Lambda^M} \left(\|\mathbf{y} - \boldsymbol{b}_{\boldsymbol{\theta}}\|^2 + \frac{1}{2} \sum_{j=1}^M \theta_j \|\boldsymbol{b}_j - \boldsymbol{b}_{\boldsymbol{\theta}}\|^2 \right).$$

This procedure satisfies the following oracle inequality [32]. For all $x > 0$, with probability greater than $1 - e^{-x}$ we have

$$\|\boldsymbol{b}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\mu}\|^2 \leq \min_{j=1,...,M} \|\boldsymbol{b}_j - \boldsymbol{\mu}\|^2 + \frac{4\sigma^2(x + \log M)}{n}.$$

Thus the estimator $\boldsymbol{b}_{\hat{\boldsymbol{\theta}}}$ mimics the best approximation of $\boldsymbol{\mu}$ among $\{\boldsymbol{b}_1, ..., \boldsymbol{b}_M\}$. This result and the procedure (1.34) first appeared in [32]. It shows that in this setting, the penalty (1.33) leads to an oracle inequality with an error term of order $\frac{\log M}{n}$. We now present similar results in other settings.

Density estimation. Here, $\mathcal{H}$ is the set of all measurable functions $\mathbb{R} \to \mathbb{R}$ and $\|g\|_{\mathcal{H}}^2 = \int_{\mathbb{R}} g(x)^2 dx$. Consider an unknown density f on $\mathbb{R}$ and let $f_1, ..., f_M$ be elements of $\mathcal{H}$. The goal is to estimate f from i.i.d. observations $X_1, ..., X_n$ drawn from f. The abstract criterion (1.34) takes the form

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Lambda^M} \left(\int_{\mathbb{R}} f_{\boldsymbol{\theta}}(x)^2 dx - \frac{2}{n} \sum_{i=1}^n f_{\boldsymbol{\theta}}(X_i) + \frac{1}{2} \sum_{j=1}^M \theta_j \int_{\mathbb{R}} (f_j(x) - f_{\boldsymbol{\theta}}(x))^2 dx \right),$$

where $f_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j f_j$. This procedure is studied in Chapter 2, where it is shown that the oracle inequality

$$\int_{\mathbb{R}} (f_{\hat{\boldsymbol{\theta}}}(x) - f(x))^2 dx \leq \min_{j=1,...,M} \int_{\mathbb{R}} (f_j(x) - f(x))^2 dx + c \left(|f|_\infty + \max_{j=1,...,M} |f_j|_\infty \right) \frac{x + \log M}{n} \quad (1.35)$$

holds with probability greater than $1 - e^{-x}$ for all $x > 0$, where $c > 0$ is an absolute constant. The oracle inequality above is the main result of Chapter 2. Several related results are included in Chapter 2.

- We prove that the tail probabilities of (1.35) are optimal in a minimax sense.
- Define $\hat{f} = \operatorname{argmin}_{g \in \{f_1, ..., f_M\}} \int_{\mathbb{R}} g(x)^2 dx - \frac{2}{n} \sum_{i=1}^n g(X_i)$ as the estimator that selects the best function in the dictionary $\{f_1, ..., f_M\}$ with respect to the empirical criterion

  $$g \to \int_{\mathbb{R}} g(x)^2 dx - \frac{2}{n} \sum_{i=1}^n g(X_i).$$

  This estimator $\hat{f}$ satisfies the oracle inequality (1.1) with $\delta_{n,M}$ of the order $(\log(M)/n)^{1/2}$. This procedure satisfies a similar oracle inequality in deviation.

- Let $\hat{k}$ be any random variable valued in $\{1, ..., M\}$ A procedure of the form $f_{\hat{k}}$ cannot achieve an oracle inequality of the form (1.1) with an error term of order $\log(M)/n$. This result is commonly known as the sub-optimality of selectors.

- The aggregate with exponential weights cannot achieve an oracle inequality of the form (1.35) with high probability.

In density estimation, our results focus on aggregation of deterministic functions since it is always possible to split the data into two independent samples, as explained in Section 1.1.1 above.

Aggregation of linear estimators in fixed design regression. We now come back to the regression model (1.13) and consider linear estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$. For all $j = 1, ..., M$ the estimator $\hat{\boldsymbol{\mu}}_j$ has the form $A_j\mathbf{y}$ for some deterministic matrix A_j. The goal is to mimic the performance of the best estimator among $\{\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M\}$. The main difference from the two previous paragraphs is that here, the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ are not deterministic as they clearly depend on the data $\mathbf{y}$.

Now that the estimators to aggregate are random and depend on the data $\mathbf{y}$, the following important questions arise.

1. Does the price to pay for aggregation increase because of the dependence between $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ and the data $\mathbf{y}$, or is it still of order $\sigma^2 \log(M/\delta)$? Is there an extra price to pay to handle the dependence?

2. A natural quantity that captures the statistical complexity of a given estimator $\hat{\boldsymbol{\mu}}_j$ is the variance defined by $\mathbb{E}\|\hat{\boldsymbol{\mu}}_j - \mathbb{E}\hat{\boldsymbol{\mu}}_j\|^2$. When the estimators are deterministic, their variances are all zero. Now that the estimators are random, does the price to pay for aggregation depend on the statistical complexities of the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$, for example through their variances? Is it harder to aggregate estimators with large statistical complexities?

Chapter 3 investigates these questions. For linear estimators, an unbiased risk estimate is given by Mallows [73] C_p criterion. For any matrix A, this criterion is defined as $C_p(A) = \|A\mathbf{y} - \mathbf{y}\|^2 + \frac{2\sigma^2 \mathrm{Tr}A}{n}$. Let $A_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j A_j$ for all $\boldsymbol{\theta} \in \mathbb{R}^M$. In this setting, the abstract procedure (1.34) takes the form

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Lambda^M} \left(C_p(A_{\boldsymbol{\theta}}) + \frac{1}{2} \sum_{j=1}^M \theta_j \|A_j\mathbf{y} - A_{\boldsymbol{\theta}}\mathbf{y}\|^2 \right). \tag{1.36}$$

The estimator $\hat{\boldsymbol{\theta}}$ is equal to the second component of (1.15). The main result of Chapter 3 is that this estimator satisfies with probability greater than $1 - e^{-x}$ the oracle inequality

$$\|A_{\hat{\boldsymbol{\theta}}}\mathbf{y} - \boldsymbol{\mu}\|^2 \leq \min_{j=1,...,M} \|A_j\mathbf{y} - \boldsymbol{\mu}\|^2 + \frac{c\sigma^2(x + \log M)}{n} \tag{1.37}$$

where $c > 0$ is an absolute constant, provided that $\max_{j=1,...,M} |||A_j|||_2 \leq 1$ where $|||\cdot|||_2$ denotes the operator norm. This result answers the two questions raised above for linear estimators in the following way.

1. The dependence between the preliminary estimators and the data $\mathbf{y}$ induces no extra cost for linear estimators. The price to pay for aggregation is still of order $\frac{\log M}{n}$.

2. The variances of the linear estimators – or any other measure of their statistical complexities – have no impact on the price to pay for aggregation.

1.2.2 Mimicking the best Lasso estimator

The estimator (1.36) enjoys the oracle inequality (1.37) in the situation where the preliminary estimators $(\hat{\boldsymbol{\mu}}_j)_{j=1,\dots,M}$ are linear estimators. If a design matrix $\mathbb{X}$ is given, we will explain in Chapter 4 that an estimator related to (1.36) can aggregate a dictionary of nonlinear estimators, provided that all nonlinear estimators $\hat{\boldsymbol{\mu}}_j$ in the dictionary are of the form $\mathbb{X}\hat{\boldsymbol{\beta}}$ for some estimator $\hat{\boldsymbol{\beta}}$ valued in $\mathbb{R}^p$. In particular, Chapter 4 suggests an estimator $\hat{\boldsymbol{\mu}}$ that satisfies

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \mathbb{E} \min_{\lambda>0} \left(\|\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda - \boldsymbol{\mu}\|^2 + \frac{c\sigma^2 |\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda|_0}{n} \log\left(\frac{ep}{|\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda|_0 \vee 1} \right) \right), \tag{1.38}$$

where $a \vee b = \max(a, b)$, $c > 0$ is a numerical constant and where $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda$ denotes the Lasso estimator with design matrix $\mathbb{X}$ and tuning parameter λ. This result holds with no assumption on the design matrix $\mathbb{X}$ or on the true mean $\boldsymbol{\mu}$.

Inequality (1.38) above is of the same nature as (1.12) in density estimation. Instead of assuming a specific structure or regularity of the design matrix or of the true parameter, the goal is to construct an estimator that is nearly as good as the best estimator in a given collection. Here, the collection of estimators is $(\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda)_{\lambda>0}$ while in Section 1.1.1, the collection of density estimators is $(f_{N,h})_{h>0}$ where $f_{N,h}$ is the kernel estimator (1.11).

Consider the following problem.

- What are suitable assumptions on the matrix $\mathbb{X}$ such that computationally tractable and consistent estimators of $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}^*$ exist in the situation $p \gg n$ and $|\boldsymbol{\beta}^*|_0 \ll n$?

This question has been investigated thoroughly in the last decade [78, 17, 61, 24]. The Lasso and the Dantzig selector are two examples of estimators that are computationally tractable and consistent in a sparse setting. Inequality (1.38) is proved in Chapter 4, it is the result of another investigation that contrasts with the above question and the results [78, 17, 61, 24]. Namely, Chapter 4 investigates the following question.

- With no assumption on the design matrix $\mathbb{X}$, is it possible to construct an estimator $\hat{\boldsymbol{\mu}}$ that is nearly as good as the best Lasso estimator? That is, is it possible to construct an estimator $\hat{\boldsymbol{\mu}}$ whose prediction performance is comparable to the benchmark $\min_{\lambda>0} \|\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda - \boldsymbol{\mu}\|^2$?

1.2.3 Advances in shape restricted regression

The second and third part of the thesis deal with shape restricted regression in the model (1.13). The two most famous examples of shape restrictions are monotonicity

and convexity. Isotonic regression is the univariate regression model where the underlying nonparametric function class is that of all nondecreasing functions, cf. (1.18). In convex regression, the underlying nonparametric function class is that of convex functions. These nonparametric classes have two remarkable properties:

- The estimation problem over these nonparametric classes admits nonparametric minimax rates similar to those obtained for smooth parametric classes such as Hölder or Sobolev ellipsoids. In the L^2 norm, the minimax rate of estimation over the class of nondecreasing functions is of order $n^{-2/3}$ while the rate of estimation over the class of convex functions is of order $n^{-4/5}$, see Chapters 5 and 6 and the references therein.

- These nonparametric classes enjoy an almost parametric phenomenon similar to sparsity phenomena in high-dimensional statistics. In high-dimensional statistics, it is shown that one can estimate an s-sparse vector in the squared ℓ_2-norm at the rate $\frac{s}{n}$ up to logarithmic factors. In isotonic regression, we show that it is possible to estimate a piecewise constant nondecreasing function at the rate $\frac{k}{n}$ if the unknown nondecreasing function has only k pieces, or equivalently $k-1$ jumps. In convex regression, it is possible to estimate a piecewise constant convex function at the rate $\frac{q}{n}$ if the unknown convex function is piecewise affine with at most q pieces, or equivalently with at most $q-1$ changes of slope. In all the cases, we deal with the squared ℓ_2-norm.

Sharp oracle inequalities through aggregation. Aggregation methods such as the procedure studied in Chapter 3 can be used to construct estimators that satisfy the almost-parametric phenomena mentioned above. Namely, Chapter 5 constructs an estimator $\hat{\boldsymbol{\mu}}$ such that with probability at least $1-e^{-x}$,

$$\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2 \le \min_{\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u}-\boldsymbol{\mu}\|^2 + \frac{c\sigma^2 k(\boldsymbol{u}) \log\left(\frac{en}{k(\boldsymbol{u})}\right)}{n} + \frac{c\sigma^2 x}{n} \right), \tag{1.39}$$

for all $x>0$ and where $c>0$ is a numerical constant. In the above display, the integer $k(\boldsymbol{u})$ is the smallest integer $k\ge 1$ such that $\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}$ is piecewise constant with k pieces. Similarly, one may define the set of convex sequences by

$$\mathcal{S}_n^{\cup} = \{\boldsymbol{u}=(u_1,...,u_n)^T\in\mathbb{R}^n : u_{i-1}+u_{i+1}\ge 2u_i, \quad i=2,...,n-1\}.$$

Here, the new quantity that characterizes the parametric phenomenon is the number of affine pieces of the sequence $\boldsymbol{u}\in\mathcal{S}_n^{\cup}$, or equivalently the number of changes of slope. Let $q(\boldsymbol{u})$ be the smallest number $q\ge 1$ such that $\boldsymbol{u}\in\mathcal{S}_n^{\cup}$ is piecewise affine with q pieces (see Chapters 5 and 6 for a precise definition of the integer $q(\boldsymbol{u})$). As for nondecreasing sequences above, in Chapter 5 we suggest an estimator $\hat{\boldsymbol{\mu}}'$ that satisfies

$$\|\hat{\boldsymbol{\mu}}'-\boldsymbol{\mu}\|^2 \le \min_{\boldsymbol{u}\in\mathcal{S}_n^{\cup}} \left(\|\boldsymbol{u}-\boldsymbol{\mu}\|^2 + \frac{c\sigma^2 q(\boldsymbol{u}) \log\left(\frac{en}{q(\boldsymbol{u})}\right)}{n} + \frac{c\sigma^2 x}{n} \right), \tag{1.40}$$

for all $x>0$ and another numerical constant $c>0$. The estimator $\hat{\boldsymbol{\mu}}$ that satisfies (1.39) and the estimator $\hat{\boldsymbol{\mu}}'$ are both constructed by aggregating projection estimators, using an aggregation procedure studied in Chapter 3. Chapter 5 thus establishes a link between shape restricted regression and the aggregation methods studied in Chapters 2 and 3.

Least Squares estimator over convex sets. Obtaining such results using aggregation methods was our initial motivation for studying shape restricted regression. Chapters 6 and 7 take a closer look at shape restricted regression. These two chapters focus on the two Least Squares estimators

$$\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) = \operatorname*{argmin}_{\boldsymbol{v}\in\mathcal{S}_n^{\uparrow}} \|\boldsymbol{v}-\mathbf{y}\|^2, \qquad \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}) = \operatorname*{argmin}_{\boldsymbol{v}\in\mathcal{S}_n^{\cup}} \|\boldsymbol{v}-\mathbf{y}\|^2.$$

We show in Chapter 6 that the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ satisfies a sharp oracle inequality similar to (1.39), while $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ satisfies a sharp oracle inequality similar to (1.40). Oracle inequalities for these two Least Squares estimators were previously known by Guntuboyina and Sen [50], Chatterjee et al. [27]. Oracle inequalities from these previous papers have a leading constant strictly greater than 1. Chapter 6 provides general techniques to obtain oracle inequalities with leading constant 1 for Least Squares estimators over convex sets.

Adaptive confidence sets: is it possible to infer that the rate of estimation is actually fast? The oracle inequality (1.39) implies that if the true regression vector $\boldsymbol{\mu}$ is nondecreasing with few jumps (i.e., $k(\boldsymbol{\mu})$ is small), then it is possible to estimate $\boldsymbol{\mu}$ at the rate $\frac{\sigma^2 k(\boldsymbol{\mu})}{n}$ up go logarithmic factors. Similarly, the oracle inequality (1.40) implies that if $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$ is convex with few changes of slope (i.e., $q(\boldsymbol{\mu})$ is small), then it is possible to estimate $\boldsymbol{\mu}$ at the rate $\frac{\sigma^2 q(\boldsymbol{\mu})}{n}$ up to logarithmic factors. The values $k(\boldsymbol{\mu})$ or $q(\boldsymbol{\mu})$ are unknown in practice, which raises the following natural question. In isotonic regression, if $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ and $k(\boldsymbol{\mu}) = k_0$, is it possible to infer from the data that the rate of estimation is at most of order $\frac{\sigma^2 k_0}{n}$ without the knowledge of k_0? This question can be answered by constructing confidence sets. We show in Chapter 7 that for isotonic and convex regression, it is possible to construct satisfactory confidence sets, so that it is possible to infer the rate of estimation from the data.

1.3 Organization of the chapters

The following chapters are self-contained and can be read independently. A short summary of each chapter and their main contribution is as follows.

- Chapter 2 studies the problem of aggregation of deterministic functions in density estimation with the L^2-loss. The main results of this chapter are the oracle inequalities given Theorems 2.6 and 2.8 and the lower bounds given in Theorems 2.1 and 2.9.

- Chapter 3 studies the problem of aggregation of affine and linear estimators in regression with fixed design. No data splitting is performed. The estimators depend on the same data as that used for aggregation. The main result of Chapter 3 is the oracle inequality given in Theorem 3.1.

- In Chapter 4, we construct an estimator that aggregates the Lasso estimators on the Lasso path. This estimator is nearly as good as the best Lasso estimator, cf. Theorem 4.3.

- Chapter 5 links the two areas of statistics studied in the thesis: Aggregation of estimators and shape restricted regression. Chapter 5 uses aggregation methods as building blocks to construct rate optimal estimators in shape restricted regression, cf. the oracle inequalities given in Theorems 5.1 and 5.6 and the lower bounds given in Propositions 5.4 and 5.7.

- Chapter 6 studies the Least Squares estimator in shape restricted regression. The main result of Chapter 6 is that the oracle inequalities obtained with aggregation methods in Chapter 5 are also satisfied for the Least Squares estimator, cf. Theorems 6.2 and 6.6.

- Finally, in Chapter 7 we construct adaptive confidence sets in the context of shape restricted regression. Chapters 5 and 6 study estimators that achieve the minimax rate of estimation on classes of monotone and convex functions. Chapter 7 proves the existence of confidence sets that capture the true function with high probability and whose diameter is of order of the minimax rate, cf. Theorems 7.2, 7.3, 7.9 and 7.11.

1.4 Bibliographic notes

The first results on aggregation in statistical settings appeared in Nemirovski [83], Catoni [25], Yang [106] and Tsybakov [98]. These early works studied three different aggregation problem.

- For the *model selection* type aggregation problem, the goal is to mimic the best function in the dictionary. Results for this problem were obtained in [106, 25, 98, 64, 70, 58, 4, 35, 90, 32, 33], Chapters 2 and 3.

- For the *convex* aggregation problem, the goal is to mimic the best convex combination of the functions in the dictionary. [98, 89, 90, 96]. Proposition 3.10 in Chapter 3 provides a convex aggregation result for affine estimators.

- For the *linear* aggregation problem, the goal is to mimic the best function in the span of the functions in the dictionary [98, 89, 90, 96].

Some more recent papers studied the *sparse* and *sparse-convex* aggregation problems, cf. [71, 92, 93, 96].

This thesis focuses on the model selection type aggregation problem. The penalized estimator studied in Chapters 2 and 3 is similar to the Q-aggregation procedure proposed by Rigollet [90] and Dai et al. [32]. Aggregation of affine estimators using the Q-aggregation procedure was previously studied in Dai et al. [33].

Aggregation of density estimators with respect to the Kullback-Leibler loss was studied in [106, 25, 58, 64]. These results hold with no boundedness assumption, unlike the aggregation problem with respect to the L^2 loss where an l_∞-bound is required on the true density to obtain satisfactory results, cf. [89, 58] and Chapter 2.

Leung and Barron [70] derived the first result on aggregation of linear estimators, where one has to handle the dependence between the estimators in the dictionary and the data used to aggregate them. These results were later generalized in Dalalyan and Salmon [35], Dai et al. [33] and in Chapter 3 of the present thesis. To our

knowledge, Corollary 4.2 in Chapter 4 is the first result on aggregation of nonlinear estimators where the nonlinear estimators are based on the same data as that used for aggregation.

The results above hold under a boundedness or subgaussian assumption. Let us mention the work of Mendelson [79] who recently studied aggregation of functions with heavy tailed noise. This result weakens the subgaussian assumption required in the papers mentioned above.

Chapter 5 explains how aggregation methods can be used to derive oracle inequalities in the context of shape restricted regression. In this context, previously obtained oracle inequalities for the Least Squares estimators can be found in Guntuboyina and Sen [50], Chatterjee et al. [27] and Chatterjee et al. [28]. These papers first studied the almost parametric phenomenon that appears if the true regression function has some low-dimensional property, cf. Chapters 5 and 6 for rigorous results and more discussion about this almost parametric phenomenon. To our knowledge, Chapter 7 provides the first results on the construction of adaptive confidence sets related to this almost parametric phenomenon.

Part I

Aggregation

Chapter 2

Optimal exponential bounds for aggregation of density estimators

We consider the problem of model selection type aggregation in the context of density estimation. We first show that empirical risk minimization is sub-optimal for this problem and it shares this property with the exponential weights aggregate, empirical risk minimization over the convex hull of the dictionary functions, and all selectors. Using a penalty inspired by recent works on the Q-aggregation procedure, we derive a sharp oracle inequality in deviation under a simple boundedness assumption and we show that the rate is optimal in a minimax sense. Unlike the procedures based on exponential weights, this estimator is fully adaptive under the uniform prior. In particular, its construction does not rely on the sup-norm of the unknown density. By providing lower bounds with exponential tails, we show that the deviation term appearing in the sharp oracle inequalities cannot be improved.

Key Words: aggregation, model selection, sharp oracle inequality, density estimation, concentration inequality, lower bounds, minimax optimality.

2.1 Introduction

We study the problem of estimation of an unknown density from observations. Let $(\mathcal{X}, \mu)$ be a measurable space. We are interested in estimating an unknown density f with respect to the measure μ given n independent observations $X_1, \dots, X_n$ drawn from f. We measure the quality of estimation of f by the L^2 squared distance

$$\|\hat{g} - f\|^2 = \int (f - \hat{g})^2 d\mu = \|\hat{g}\|^2 - 2\int \hat{g} f d\mu + \|f\|^2, \tag{2.1}$$

for any $\hat{g} \in L^2(\mu)$ possibly dependent on the data $X_1, \dots, X_n$. Since the term $\|f\|^2$ is constant for all $\hat{g}$, we will consider throughout the paper the risk

$$R(\hat{g}) = \|\hat{g}\|^2 - 2\int \hat{g} f d\mu. \tag{2.2}$$

An estimator $\hat{g}$ minimizes $R(\cdot)$ if and only if it minimizes (2.1).

Given M functions $f_1, \dots, f_M \in L^2(\mu)$, we would like to construct a measurable function $\hat{g}$ of the observations $X_1, \dots, X_n$ that is almost as good as the best function among $f_1, \dots, f_M$. The model may be misspecified, which means that f may not be one of the functions $f_1, \dots, f_M$. We are interested in deriving oracle inequalities,

either in expectation

$$\mathbb{E}R(\hat{g}) \leq C \min_{j=1,\dots,M} R(f_j) + \delta_{n,M},$$

or with high probability, i.e., for all $\varepsilon > 0$, with probability greater than $1 - \varepsilon$

$$R(\hat{g}) \leq C \min_{j=1,\dots,M} R(f_j) + \delta_{n,M} + d_{n,M}(\varepsilon),$$

where $\delta_{n,M}$ is a small quantity and $d_{n,M}(\cdot)$ is a function of ε that we call the deviation term. We are only interested in sharp oracle inequalities, i.e., oracle inequalities where the leading constant is $C = 1$, since it is essential to derive minimax optimality results.

We consider only deterministic functions for $f_1, \dots, f_M$. They cannot depend on the data $X_1, \dots, X_n$. A standard application of this setting was introduced in Wegkamp [104]: given $m + n$ i.i.d. observations drawn from f, use the first m observations to build M estimators $\hat{f}_1, \dots, \hat{f}_M$, and in a second step use the remaining n observations to select the best among the preliminary estimators $\hat{f}_1, \dots, \hat{f}_M$. A related problem is selecting the best estimator from a family $\hat{f}_1, \dots, \hat{f}_M$ where these estimators are built using the same data used for model selection or aggregation. Such problems were recently considered in Dalalyan and Salmon [35] and Dai et al. [33] for the regression model with fixed design.

We are also interested in deriving sharp oracle inequalities with prior weights on the model $\{f_1, \dots, f_M\}$. To be more precise, for some prior probability distribution $\pi_1, \dots, \pi_M$ over the finite set $\{f_1, \dots, f_M\}$ and any $\varepsilon > 0$, our estimator $\hat{f}_n$ should satisfy with probability greater than $1 - \varepsilon$

$$R(\hat{f}_n) \leq \min_{j=1,\dots,M} \left(R(f_j) + \frac{\beta}{n} \log \frac{1}{\pi_j} \right) + d_{n,M}(\varepsilon), \tag{2.3}$$

for some positive constant β and some deviation term $d_{n,M}(\cdot)$. The Mirror Averaging algorithm [58, 37] is known to achieve a similar oracle inequality in expectation. The analysis of Juditsky et al. [58] shows that the constant β scales linearly with the sup-norm of the unknown density, which is also the case for the results presented here. Model selection techniques with prior weights were used in order to derive sparsity oracle inequalities using sparsity pattern aggregation [92, 93, 37].

Another related learning problem is that of model selection when the model is finite dimensional with a specific shape, for example a linear span of M functions or the convex hull of M functions. This is the aggregation framework and it has received a lot of attention in the last decade to construct adaptive estimators that achieve the minimax optimal rates, especially for the regression problem [98, 71, 92, 65, 93] but also for density estimation [106, 64, 89].

The main contribution of the present paper is the following.

- We provide sharp oracle inequalities and the corresponding tight lower bounds for two procedures: empirical risk minimization over the discrete set $\{f_1, \dots, f_M\}$ and the penalized procedure (2.14) with the penalty (2.15). Here, tight means that neither the rate nor the deviation term of the sharp oracle inequalities can be improved. The sharp oracle inequalities are given in Theorem 2.2 and Corollary 2.8 and the tight lower bounds are given in Theorem 2.1 and Theorem 2.9. These results lead to a definition of minimax optimality in deviation, which is discussed in Section 2.4.

While proving the above results, we extend several aggregation results that are known for the regression model to the density estimation setting. Let us relate these results of the present paper to the existing literature on the regression model:

- In Theorem 2.2, we derive a sharp oracle inequality in deviation for the empirical risk minimizer over the discrete set $\{f_1, \dots, f_M\}$. This is new in the context of density estimation, and an analogous result is known for the regression model [93].

- In Theorem 2.6, we derive a sharp oracle inequality in deviation for penalized empirical risk minimization with the penalty (2.15). With the uniform prior, this yields the correct rate $(\log M)/n$ of model selection type aggregation. This penalty is inspired by recent works on the Q-aggregation procedure [69, 32] where similar oracle inequalities in deviation were obtained for the regression model. The first sharp oracle inequalities that achieve the correct rate of model selection type aggregation were obtained in expectation for the regression model in [106, 25].

- We extend several lower bounds known for the regression model to the density estimation setting. We show that any procedure that selects a dictionary function cannot achieve a better rate than $\sqrt{(\log M)/n}$ and that the rate of model selection type aggregation is of order $(\log M)/n$. We also show that the exponential weights aggregate and the empirical risk minimizer over the convex hull of the dictionary functions cannot be optimal in deviation, with an unavoidable error term of order $1/\sqrt{n}$. Earlier results for the regression model can be found in [98, 93] for lower bounds on model selection type aggregation and the performance of selectors, while [66, 32, 68] contain earlier lower bounds on the performance of exponential weights and empirical risk minimization over the convex hull of the dictionary.

An aspect of our results is not present in the previous works on the regression model. In the literature on aggregation in the regression model, lower bounds are proved either in expectation or in probability in the form

$$\mathbb{P}\left(R(\hat{T}_n) > \min_{j=1,\dots,M} R(f_j) + \psi_{n,M}\right) > c, \tag{2.4}$$

for any estimator $\hat{T}_n$, a risk function $R(\cdot)$, a rate $\psi_{n,M}$ and some absolute constant $c > 0$, usually $c = 1/2$. The tight lower bounds presented in Theorem 2.1 and Theorem 2.9 contrast with lower bounds of the form (2.4) as they yield for any estimator $\hat{T}_n$,

$$\forall x > 0, \quad \mathbb{P}\left(R(\hat{T}_n) > \min_{j=1,\dots,M} R(f_j) + \psi_{n,M} + \frac{x}{n}\right) > c\exp(-x), \tag{2.5}$$

i.e., they provide lower bounds for any probability estimate in an interval $(0, 1/c)$ where $c > 0$ is an absolute constant. Moreover, these lower bounds show that the exponential tail of of the excess risk of the estimators from Theorem 2.2 and Theorem 2.6 cannot be improved. The tools used in the present paper to prove lower bounds of the form (2.5), in particular Lemma 2.16, can be used to prove similar results for regression model. The tight lower bounds of the present paper contrast

with the existing literature on the regression model, since to our knowledge, there is no lower bound of the form (2.5) available for regression.

In the regression model with random design, given a class of functions G, a penalty $\mathrm{pen}(\cdot)$, a coefficient $\nu > 0$ and observations $(X_1, Y_1), ..., (X_n, Y_n)$, penalized empirical risk minimization solves the optimization problem

$$\min_{g \in G} \quad \frac{1}{n} \sum_{i=1}^{n} (g(X_i) - Y_i)^2 + \nu \mathrm{pen}(g). \tag{2.6}$$

But if the distribution of the design is known, the statistician can compute the quantity $\mathbb{E}[g(X)^2]$ for all $g \in G$ and solve the following minimization problem that slightly differs from (2.6):

$$\min_{g \in G} \quad \mathbb{E}[g(X)^2] - \frac{2}{n} \sum_{i=1}^{n} g(X_i) Y_i + \nu \mathrm{pen}(g). \tag{2.7}$$

In the regression model, the distribution of the design is rarely known so the penalized ERM that solves (2.7) has not received as much attention as the procedure (2.6) when the distribution of the design is not known. The density estimation setting studied in the present paper is closer to the regression setting with known design (2.7) than to the regression setting with unknown design (2.6) studied in [69]. There are differences with respect to the choice of coefficient of the penalty (2.15), and to the form of the empirical process that appears in the analysis. These differences are more thoroughly discussed in Section 2.3.4.

The paper is organized as follows. In Section 2.2 we show that empirical risk minimization achieves a sharp oracle inequality with slow rate, but this rate cannot be improved among selectors. Two classical estimators, the exponential weights aggregate and empirical risk minimization over the convex hull of the dictionary functions, are shown to be suboptimal in deviation. In Section 2.3, we define a penalized procedure that achieves the optimal rate $\frac{\log M}{n}$ in deviation, and we provide a lower bound that shows that neither the rate nor the deviation term can be improved. Section 2.4 proposes a definition of minimax optimality in deviation and shows that it is satisfied by the procedures given in Sections 2.2 and 2.3. Section 2.5 is devoted to the proofs.

2.2 Sub-optimality of selectors, ERM and exponential weights

2.2.1 Selectors

Define a selector as a function of the form $f_{\hat{J}}$ where $\hat{J}$ is measurable with respect to $X_1, \ldots, X_n$ with values in $\{1, \ldots, M\}$. It was shown in the regression framework [58, 93] that selectors are suboptimal and cannot achieve a better rate that $\sigma\sqrt{\frac{\log M}{n}}$ where σ^2 is the variance of the regression noise. The following theorem extends this lower bound for selectors to density estimation. The underlying measure μ is the Lebesgue measure on $\mathbf{R}^d$ for $d \geq 1$.

Theorem 2.1 (Lower bounds for selectors)**.** *Let $L > 0$, and $M \geq 2, n \geq 1, d \geq 1$ be integers. Let $\mathcal{F}$ be the class of all densities f with respect to the Lebesgue measure*

on $\mathbf{R}^d$ *such that* $\|f\|_\infty \le L$. *Let* $x \ge 0$ *satisfying*

$$\frac{\log(M) + x}{n} < 3.$$

Then there exist $f_1, \dots, f_M \in L^2(\mathbf{R}^d)$ *with* $\|f_j\|_\infty \le L$ *such that the following lower bound holds:*

$$\inf_{\hat{S}_n} \sup_{f \in \mathcal{F}} \mathbb{P}_f \left(\|\hat{S}_n - f\|^2 - \inf_{j=1,\dots,M} \|f_j - f\|^2 \ge \frac{L}{\sqrt{3}} \sqrt{\frac{x + \log M}{n}} \right) \ge \frac{1}{24} \exp(-x)$$

where $\mathbb{P}_f$ *denotes the probability with respect to* n *i.i.d. observations with density* f *and the infimum is taken over all selectors* $\hat{S}_n$.

The proof of Theorem 2.1 is given in Section 2.5. It can be extended to other measures as soon as the underlying measurable space allows the construction of an orthogonal system such as the one described in Proposition 2.15 below.

For any $g \in L^2(\mu)$, define the empirical risk

$$R_n(g) = \|g\|^2 - \frac{2}{n} \sum_{j=1}^{M} g(X_i). \tag{2.8}$$

The empirical risk (2.8) is an unbiased estimator of the risk (2.2). In order to explain the idea behind the proof of our main result described in Theorem 2.6, it is useful the prove the following oracle inequality for the empirical risk minimizer over the discrete set $\{f_1, \dots, f_M\}$.

Theorem 2.2. *Assume that the functions* $f_1, \dots, f_M \in L^2(\mu)$ *satisfy* $|f_j|_\infty \le L_0$ *for all* $j = 1, \dots, M$. *Define*

$$\hat{J} \in \operatorname*{argmin}_{j=1,\dots,M} \left(\|f_j\|^2 - \frac{2}{n} \sum_{i=1}^{n} f_j(X_i) \right).$$

Then for any $x > 0$, *with probability greater than* $1 - \exp(-x)$,

$$R(f_{\hat{J}}) \le \min_{j=1,\dots,M} R(f_j) + L_0 \left(4\sqrt{2} \sqrt{\frac{x + \log M}{n}} + \frac{8(x + \log M)}{3n} \right).$$

Together with Theorem 2.1, Theorem 2.2 shows that empirical risk minimization is optimal among selectors. Unlike the oracle inequality of Theorem 2.6 below, this result applies for any density f, with possibly $|f|_\infty = \infty$. Its proof relies on the concentration of $R_n(g) - R(g)$ around 0 for fixed functions g with $|g|_\infty \le L_0$.

Proof of Theorem 2.2. We will use the following notation that is common in the literature on empirical processes. For any $g \in L^2(\mu)$, define

$$\begin{aligned} Pg &= \int g f d\mu, \\ P_n g &= \frac{1}{n} \sum_{i=1}^{n} g(X_i). \end{aligned} \tag{2.9}$$

With this notation, the difference between the real risk (2.2) and the empirical risk (2.8) can be rewritten

$$R(g) - R_n(g) = (P - P_n)(-2g). \tag{2.10}$$

Let J^* be such that $R(f_{J^*}) = \min_{j=1,\dots,M} R(f_j)$. The definition of $\hat{J}$ yields $R_n(f_{\hat{J}}) \leq R_n(f_{J^*})$. Using (2.10), it can be rewritten

$$R(f_{\hat{J}}) - R(f_{J^*}) \leq (P - P_n)(-2f_{\hat{J}} + 2f_{J^*}).$$

We can control the right hand side of the last display using the concentration inequality (2.23) with a union bound over $j = 1, \dots, M$. For any $t > 0$, with probability greater than $1 - M\exp(-t)$,

$$\begin{aligned}(P - P_n)(-2f_{\hat{J}} + 2f_{J^*}) &\leq \max_{j=1,\dots,M} (P - P_n)(-2f_j + 2f_{J^*}), \\ &\leq \sigma\sqrt{\frac{2t}{n}} + \frac{8L_0 t}{3n},\end{aligned}$$

where $\sigma^2 = \max_{j=1,\dots,M} P(-2f_j + 2f_{J^*})^2 \leq 16L_0^2$. Setting $x = t - \log M$ yields the desired oracle inequality. □

By inspecting the short proof above, we see that the slow rate term $\sqrt{\frac{x+\log M}{n}}$ comes from the variance term in the concentration inequality (2.23).

We can draw two conclusions from Theorems 2.1 and 2.2.

- In order to achieve faster rates than $\sqrt{\frac{\log M}{n}}$, we need to look for estimators taking values beyond the discrete set $\{f_1, \dots, f_M\}$. In Section 2.3, we will consider estimators taking values in the convex hull of this discrete set.
- The proof of Theorem 2.2 suggests that a possible way to derive an oracle inequality with fast rates is to cancel the variance term in the concentration inequality (2.23). In order to do this, we need some positive gain on the empirical risk of our estimator. Namely, for some oracle j^* we would like our estimator $\hat{f}_n$ to satisfy $R_n(\hat{f}_n) \leq R_n(f_{J^*})$ minus some positive value. This value is given by the strong convexity of the empirical objective in Proposition 2.7.

Define the simplex in $\mathbb{R}^M$:

$$\Lambda^M = \left\{ \theta \in \mathbb{R}^M, \quad \sum_{j=1}^M \theta_j = 1, \quad \forall j = 1 \dots M, \theta_j \geq 0 \right\}. \tag{2.11}$$

Given a finite set or *dictionary* $\{f_1, \dots, f_M\}$, define for any $\theta \in \Lambda^M$

$$f_\theta = \sum_{j=1}^M \theta_j f_j. \tag{2.12}$$

In particular, $f_j = f_{e_j}$ where $e_1, \dots, e_M$ are the vectors of the canonical basis in $\mathbb{R}^M$.

Two classical estimators, the ERM over the convex hull of $f_1, ..., f_M$ and the exponential weights aggregate, are known to be sub-optimal in the regression setting [32, 66, 67, 68]. In the following we show that the same conclusions hold for density estimation with the L^2 risk.

2.2.2 ERM over the convex hull

A first natural estimator valued in the convex hull of the dictionary functions is the ERM. However, as in the regression setting [66], this estimator is suboptimal with an unavoidable error term or order $1/\sqrt{n}$.

Proposition 2.3. *Let $\mathcal{X} = \mathbf{R}$ and μ be the Lebesgue measure on $\mathbf{R}$. There exist absolute constants $C_0, C_1, C_2, C_3 > 0$ such that the following holds. Let $L > 0$. For any integer $n \geq 1$, there exist a density f bounded by L and a dictionary $\{f_1, ..., f_M\}$ of functions bounded by $2L$, with $C_0\sqrt{n} \leq M \leq C_1\sqrt{n}$, such that with probability greater than $1 - 12\exp(-C_2 M)$,*

$$\|f_{\hat{\theta}^{ERM}} - f\|^2 \geq \min_{j=1,...,M} \|f_j - f\|^2 + \frac{C_3 L}{\sqrt{n}},$$

where $\hat{\theta}^{ERM} := \operatorname{argmin}_{\theta \in \Lambda^M} R_n(f_\theta)$.

The proof of Proposition 2.3 can be found in Section 2.5.5.2.

2.2.3 Exponential Weights

The exponential weights aggregate is known to achieve optimal oracle inequalities in expectation when the temperature parameter $\beta > 0$ is chosen carefully [70, 36, 58]. Given prior weights $(\pi_1, ..., \pi_M)^T \in \Lambda^M$, it can be defined as follows:

$$\hat{f}_\beta^{EW} = \sum_{j=1}^{M} \hat{\theta}_j^{EW,\beta} f_j, \qquad \hat{\theta}^{EW,\beta} \in \Lambda^M, \qquad \hat{\theta}_j^{EW,\beta} \propto \pi_j \exp\left(-\frac{n}{\beta} R_n(f_j)\right).$$

The following proposition shows that it is suboptimal in deviation for any temperature, with a error term of order at least $1/\sqrt{n}$. This phenomenon was observed in the regression setting [32, 66], and Proposition 2.4 shows that it also holds for density estimation. As opposed to [32], the following lower bound requires only 3 dictionary functions.

Proposition 2.4. *There exist absolute constants $C_0, C_1, N_0 > 0$ such that the following holds. Let $\mathcal{X} = \mathbf{R}$ and μ be the Lebesgue measure on $\mathbf{R}$. For all $n \geq N_0, L > 0$, there exist a probability density f with respect to μ, a dictionary $\{f_1, f_2, f_3\}$ and prior weights $(\pi_1, \pi_2, \pi_3) \in \Lambda^3$ such that with probability greater than C_0,*

$$\left\|\hat{f}_\beta^{EW} - f\right\|^2 \geq \min_{j=1,2,3} \|f_j - f\|^2 + \frac{C_1 L}{\sqrt{n}},$$

Furthermore, $|f|_\infty \leq L$, and $|f_j|_\infty \leq 3L$ for $j = 1, 2, 3$.

The following proposition shows that the optimality in expectation cannot hold if the temperature is below a constant, extending a result from [66] to the density estimation setting.

Proposition 2.5. *Let $\mathcal{X} = \mathbf{R}$ and μ be the Lebesgue measure on $\mathbf{R}$. There exist absolute constants $c_0, c_1, c_2 > 0$ such that the following holds. Let $L > 0$. For any odd integer $n \geq c_0$, there exist a probability density f with respect to μ with $|f|_\infty \leq L$, and a dictionary $\{f_1, f_2\}$ with $f_j : \mathcal{X} \to \mathbf{R}$ and $|f_j|_\infty \leq L$ for $j = 1, 2$ for which the following holds:*

$$\mathbb{E}\left\|\hat{f}_\beta^{EW} - f\right\|^2 \geq \min_{j=1,2} \|f_j - f\|^2 + \frac{c_2 L}{\sqrt{n}} \text{ if } \beta \leq c_1 L.$$

The proofs of Proposition 2.4 and Proposition 2.5 can be found in Section 2.5.5.3.

2.3 Optimal exponential bounds for a penalized procedure

2.3.1 From strong convexity to a sharp oracle inequality

In this section we derive a sharp oracle inequality for the estimator $f_{\hat{\theta}}$ where $\hat{\theta}$ is defined in (2.14). Define the empirical objective H_n and the estimator $\hat{\theta}$ by

$$H_n(\theta) = \left(\|f_\theta\|^2 - \frac{2}{n}\sum_{i=1}^{n} f_\theta(X_i) \right) + \frac{1}{2}\mathrm{pen}(\theta) + \frac{\beta}{n}\sum_{j=1}^{M} \theta_j \log\frac{1}{\pi_j}, \tag{2.13}$$

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in\Lambda^M} H_n(\theta), \tag{2.14}$$

for some positive constant β and

$$\forall \theta \in \Lambda^M, \quad \mathrm{pen}(\theta) = \sum_{j=1}^{M} \theta_j \|f_\theta - f_j\|^2 . \tag{2.15}$$

The simplex Λ^M and f_θ are defined in (2.11) and (2.12).

The term

$$\frac{\beta}{n}\sum_{j=1}^{M} \theta_j \log\frac{1}{\pi_j}$$

is a penalty that assigns different weights to the functions f_j according to some prior knowledge given by $\pi_1, ..., \pi_M$, in order to achieve an oracle inequality such as (2.3).

The penalty (2.15) as well as the present procedure are inspired by recent works on Q-aggregation in regression models [90, 32, 69]. The choice of the coefficient $\frac{1}{2}$ for the penalty (2.15) is explained in Remark 2.1 below. An intuitive interpretation of the penalty (2.15) can be as follows. A point f_θ is in the convex hull of $\{f_1, \dots, f_M\}$ if and only if it is the expectation of a random variable taking values in $\{f_1, \dots, f_M\}$. The penalty (2.15) can be seen as the variance of such a random variable whose distribution is given by θ. More precisely, let η be a random variable with $\mathbb{P}(\eta = j) = \theta_j$ for all $j = 1, \dots, M$. Denote by $\mathbb{E}_\theta$ the expectation with respect to the random variable η. Then $\mathbb{E}_\theta[f_\eta] = f_\theta$ and

$$\mathrm{pen}(\theta) = \mathbb{E}_\theta \|f_\eta - \mathbb{E}_\theta[f_\eta]\|^2 ,$$

which is the variance of the random point f_η. The penalty (2.15) vanishes at the extreme points:

$$\forall j = 1, \dots, M, \qquad \mathrm{pen}(e_j) = 0,$$

and $\mathrm{pen}(\theta)$ increases as θ moves away from an extreme point e_j. Thus we convexify the optimization problem over the discrete set $\{f_1, \dots, f_M\}$ by considering the convex set $\{\mathbb{E}_\theta[f_\eta], \theta \in \Lambda^M\}$ which is exactly the convex hull of $\{f_1, \dots, f_M\}$, and we penalize by the variance of the random point f_η.

It is also possible to describe the level sets of the penalty (2.15). Assume only in this paragraph that the Gram matrix of $f_1, ..., f_M$ is invertible and let $c \in L^2(\mu)$ be in the linear span of $f_1, ..., f_M$ such that for all $j = 1, ..., M$, $\int 2cf_j d\mu = \|f_j\|^2$. Then simple algebra yields

$$\mathrm{pen}(\theta) = \|c\|^2 - \|c - f_\theta\|^2 .$$

Thus the level sets of the penalty (2.15) are euclidean balls centered at c.

Last, note that $f_{\hat{\theta}}$ coincides with the Q-aggregation procedure from [32] since

$$\left(\|f_\theta\|^2 - \frac{2}{n}\sum_{i=1}^n f_\theta(X_i)\right) + \frac{1}{2}\mathrm{pen}(\theta) = R_n(\theta) + \frac{1}{2}\mathrm{pen}(\theta) = \frac{1}{2}\left(R_n(\theta) + \sum_{j=1}^M \theta_j R_n(f_j)\right).$$

We propose an estimator $f_{\hat{\theta}}$ based on penalized empirical risk minimization over the simplex, with $\hat{\theta}$ defined in (2.14). This estimator satisfies the following oracle inequality.

Theorem 2.6. *Assume that the functions $f_1, \dots f_M$ satisfy $|f_j|_\infty \leq L_0$ for all $j = 1, \dots, M$, and assume that the unknown density f satisfies $|f|_\infty \leq L$. Let $\hat{\theta}$ be defined in (2.14) with*

$$\beta = 4L + \frac{8L_0}{3}.$$

Then for any $x > 0$, with probability greater than $1 - \exp(-x)$,

$$R(f_{\hat{\theta}}) \leq \min_{j=1,\dots,M}\left(R(f_j) + \frac{\beta}{n}\log\frac{1}{\pi_j}\right) + \frac{\beta x}{n}. \tag{2.16}$$

The following proposition specifies the property of strong convexity of the objective function $H_n(\cdot)$ defined in (2.13), which is key to prove Theorem 2.6.

Proposition 2.7 (Strong convexity of H_n)**.** *Let H_n and $\hat{\theta}$ be defined by (2.13) and (2.14), respectively. Then for any $\theta \in \Lambda^M$,*

$$H_n(\hat{\theta}) \leq H_n(\theta) - \frac{1}{2}\|f_\theta - f_{\hat{\theta}}\|^2. \tag{2.17}$$

For any $\theta \in \Lambda^M$, empirical risk minimization only grants the simple inequality

$$R_n(\hat{\theta}) \leq R_n(\theta),$$

but with Proposition 2.7 we gain the extra term $\frac{1}{2}\|f_\theta - f_{\hat{\theta}}\|^2$. To prove Theorem 2.6, we will use this extra term to compensate the variance term of the concentration inequality (2.24). Strong convexity plays an important role in our proofs, and we believe that our arguments would not work for loss functions that are not strongly convex such as the Hellinger distance, the Total Variation distance or the Kullback-Leibler divergence.

The proof of Proposition 2.7 is given in Section 2.5.3. We now give the proof of our main result, which is close to the proof of Theorem 2.2 except that we leverage the strong convexity of the empirical objective H_n.

Proof of Theorem 2.6. Note that $\mathrm{pen}(e_j) = 0$ for $j = 1, \dots, M$ and let

$$j^* \in \operatorname*{argmin}_{j=1,\dots,M}\left(\|f_j\|^2 - 2\int f_j f d\mu + \frac{\beta}{n}\log\frac{1}{\pi_j}\right) = \operatorname*{argmin}_{j=1,\dots,M}\mathbb{E}\left[H_n(e_j)\right].$$

Using (2.17) of Proposition 2.7

$$\begin{aligned}
H_n(\hat{\theta}) - H_n(e_{j^*}) &\leq -\frac{1}{2}\|f_{j^*} - f_{\hat{\theta}}\|^2, \\
R_n(\hat{\theta}) + \frac{\beta}{n}\sum_{j=1}^M \hat{\theta}_j \log\frac{1}{\pi_j} - R_n(e_{j^*}) - \frac{\beta}{n}\log\frac{1}{\pi_{j^*}} &\leq -\frac{1}{2}\|f_{j^*} - f_{\hat{\theta}}\|^2 - \frac{1}{2}\mathrm{pen}(\hat{\theta}), \\
&= -\frac{1}{2}\sum_{j=1}^M \hat{\theta}_j \|f_j - f_{j^*}\|^2,
\end{aligned}$$

where we used Proposition 2.12 with $g = f_{j^*}$ for the last display. Using (2.10), we get

$$R(f_{\hat{\theta}}) - R(f_{j^*}) - \frac{\beta}{n} \log \frac{1}{\pi_{j^*}} \leq Z_n$$

where

$$Z_n = (P - P_n)(-2f_{\hat{\theta}} + 2f_{j^*}) - \frac{\beta}{n} \sum_{j=1}^{M} \hat{\theta}_j \log \frac{1}{\pi_j} - \frac{1}{2} \sum_{j=1}^{M} \hat{\theta}_j \|f_j - f_{j^*}\|^2$$

and the notation P and P_n is defined in (2.9) and (2.10). The quantity Z_n is affine in θ and an affine function over the simplex is maximized at a vertex, so almost surely,

$$\begin{aligned} Z_n &\leq \max_{\theta \in \Lambda^M} \left(-2(P - P_n)(f_\theta - f_{j^*}) - \frac{1}{2} \sum_{j=1}^{M} \theta_j \|f_{j^*} - f_j\|^2 - \frac{\beta}{n} \sum_{j=1}^{M} \theta_j \log \frac{1}{\pi_j} \right), \\ &= \max_{k=1,\dots,M} \left(-2(P - P_n)(f_k - f_{j^*}) - \frac{1}{2} \|f_k - f_{j^*}\|^2 - \frac{\beta}{n} \log \frac{1}{\pi_k} \right). \end{aligned} \tag{2.18}$$

Let $k = 1, \dots, M$ fixed. Applying Proposition 2.14 with $g = -2(f_k - f_{j^*})$ and $\pi = \pi_k$ yields

$$\mathbb{P}\left(-2(P - P_n)(f_k - f_{j^*}) - \frac{1}{2} \|f_k - f_{j^*}\|^2 - \frac{\beta}{n} \log \frac{1}{\pi_k} > \frac{\beta x}{n} \right) \leq \pi_k \exp(-x).$$

To complete the proof, we use a union bound on $k = 1, \dots, M$ together with $\sum_{j=1}^{M} \pi_j = 1$ and (2.18):

$$\mathbb{P}\left(Z_n > \frac{\beta x}{n} \right) \leq \sum_{k=1}^{M} \pi_k \exp(-x) = \exp(-x).$$

□

Remark 2.1 (Choice of the coefficient of the penalty (2.15)). Let $\nu \in (0, 1)$. With minor modifications to the proof of Theorem 2.6, it can be shown that the oracle inequality (2.16) still holds with

$$\begin{aligned} \beta &= \frac{2L}{\min(\nu, 1 - \nu)} + \frac{8L_0}{3}, \\ H_n(\theta) &= \left(\|f_\theta\|^2 - \frac{2}{n} \sum_{i=1}^{n} f_\theta(X_i) \right) + \nu \mathrm{pen}(\theta) + \frac{\beta}{n} \sum_{j=1}^{M} \theta_j \log \frac{1}{\pi_j}, \\ \hat{\theta} &\in \operatorname*{argmin}_{\theta \in \Lambda^M} H_n(\theta). \end{aligned}$$

The oracle inequality (2.16) is best when β is small. Thus the choice $\nu = \frac{1}{2}$ is natural since it minimizes the value of β.

The optimization problem (2.14) is a quadratic program, for which efficient algorithms exist. We refer to [32, Section 4] for an analysis of the statistical performance of an algorithm that approximately solves a optimization problem similar to (2.14) in the regression setting.

The estimator $\hat{\theta}$ of Theorem 2.6 is not adaptive since its construction relies on L, an upper bound of the sup-norm of the unknown density. However, in the case

of the uniform prior $\pi_j = 1/M$ for all $j = 1, \dots, M$, Corollary 2.8 below provides an estimator which is fully adaptive: its construction depends only on the functions $f_1, \dots, f_M$ and the data $X_1, \dots, X_n$. A similar adaptivity property was observed in [69] in the regression setting.

Corollary 2.8 (Adaptive estimator)**.** *Assume that the functions $f_1, \dots f_M$ satisfy $|f_j|_\infty \leq L_0$ for all $j = 1, \dots, M$, and assume that the unknown density f satisfies $|f|_\infty \leq L$. Let*

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \Lambda^M} \left(\|f_\theta\|^2 - \frac{2}{n} \sum_{i=1}^{n} f_\theta(X_i) \right) + \frac{1}{2} \mathrm{pen}(\theta). \tag{2.19}$$

Then for any $x > 0$, with probability greater than $1 - \exp(-x)$,

$$R(f_{\hat{\theta}}) \leq \min_{j=1,\dots,M} R(f_j) + \left(4L + \frac{8L_0}{3} \right) \frac{\log(M) + x}{n}.$$

Proof of Corollary 2.8. With the uniform prior, $\pi_j = 1/M$ for all $j = 1, \dots, M$, the quantity

$$\frac{\beta}{n} \sum_{j=1}^{M} \theta_j \log \frac{1}{\pi_j} = \frac{\beta}{n} \log M$$

is independent of $\theta \in \Lambda^M$. The minimizer (2.19) is also a minimizer of the empirical objective (2.13) used in Theorem 2.6. Thus, the estimator $f_{\hat{\theta}}$ satisfies (2.16) which completes the proof. □

Corollary 2.8 is in contrast to methods related to exponential weights such as the mirror averaging algorithm from [58] as these methods rely on the knowledge of the sup-norm of the unknown density. The method presented here is an improvement in two aspects. First, the estimator of Corollary 2.8 is fully data-driven. Second, the sharp oracle inequality is satisfied not only in expectation, but also in deviation.

However, the method of Theorem 2.6 loses this adaptivity property when a non-uniform prior is used, and we do not know if it is possible to build an optimal and fully adaptive estimator for non-uniform priors.

2.3.2 A lower bound with exponential tails

The following lower bound shows that the sharp oracle inequality of Corollary 2.8 cannot be improved both in the rate and in the tail of the deviation.

Theorem 2.9 (Lower bounds with optimal deviation term)**.** *Let $M \geq 2, n \geq 1$ be two integers and let a real number $x \geq 0$ satisfy*

$$\frac{\log(M) + x}{n} < 3.$$

Let $L > 0$ and $d \geq 1$. Let $\mathcal{F}$ be the class of densities f with respect to the Lebesgue measure on $\mathbf{R}^d$ such that $\|f\|_\infty \leq L$.

Then there exist M functions $f_1, \dots, f_M$ in $L^2(\mathbf{R}^d)$ with $|f_j|_\infty \leq L$ satisfying

$$\inf_{\hat{T}_n} \sup_{f \in \mathcal{F}} \mathbb{P}_f \left(\left\| \hat{T}_n - f \right\|^2 - \min_{j=1,\dots,M} \|f_j - f\|^2 > \frac{L}{24} \left(\frac{\log(M) + x}{n} \right) \right) \geq \frac{1}{24} \exp(-x)$$

where the infimum is taken over all estimators $\hat{T}_n$ and $\mathbb{P}_f$ denotes the probability with respect to n i.i.d. observations with density f.

Notice that the restriction $\frac{\log(M)+x}{n} < 3$ is natural since the estimator $\hat{T}_n^* \equiv 0$ achieves a constant error term and is optimal in the region $\frac{\log(M)+x}{n} > c$ for some absolute constant c. Indeed, as the unknown density satisfies $|f|_\infty \leq L$, we have with probability 1:

$$\left\|\hat{T}_n^* - f\right\|^2 = \|f\|^2 \leq L \leq \inf_{j=1,\dots,M} \|f - f_j\|^2 + L, \tag{2.20}$$
$$R(\hat{T}_n^*) \leq \inf_{j=1,\dots,M} R(f_j) + L.$$

Thus it is impossible to get the lower bound of Theorem 2.9 for arbitrarily large $\frac{x+\log M}{n}$.

2.3.3 Weighted loss and unboundedness

The previous strategy based on penalized risk minimization over the simplex can be applied to handle unbounded densities or unbounded dictionary functions, if we use a weighted loss.

Let $w : \mathcal{X} \to \mathbf{R}^+$ be a measurable function with respect to μ. Define the norm (or semi-norm if w is zero on a set of positive measure)

$$\|g\|_w^2 = \int g^2 w d\mu, \qquad \forall g \in L^2(\mu).$$

Then we can define the estimator $f_{\hat{\theta}}$ where

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Lambda^M} V_n(\theta), \qquad V_n(\theta) = P_n\left(\|f_\theta\|_w^2 - \frac{2}{n}\sum_{i=1}^n f_\theta(X_i)w(X_i) + \frac{1}{2}\sum_{j=1}^M \theta_j \|f_j - f_\theta\|_w^2\right).$$

The function V_n is strongly convex with respect to the new norm $\|\cdot\|_w^2$. As in the proof of Theorem 2.6,this leads to

$$\|f_{\hat{\theta}} - f\|_w^2 \leq \|f_{j^*} - f\|_w^2 + \max_{k=1,\dots,M} \delta_k, \qquad \delta_k := (P - P_n)(-2(f_{j^*} - f_k)w) - \frac{1}{2}\|f_{j^*} - f_k\|_w^2.$$

If for some $L, L_0 > 0$, $|wf|_\infty \leq L$ and $\max_{j=1,\dots,M} |wf_j|_\infty \leq L_0$, then

$$\delta_k \leq -2(P - P_n)((f_k - f_{j^*})w) - \frac{1}{2L}\mathbb{E}[(f_k(X) - f_{j^*}(X))^2 w(X)^2].$$

We apply (2.24) to the random variables $(f_k - f_{j^*})(X_i)w(X_i)$, which are almost surely bounded by L_0. Using the union bound on $k = 1, ..., M$ we obtain $\max_{k=1,\dots,M} \delta_k \leq \beta(x + \log M)/n$ with probability greater than $1 - \exp(-x)$. and thus

$$\|f_{\hat{\theta}} - f\|_w^2 \leq \|f_{j^*} - f\|_w^2 + \beta\left(\frac{x + \log M}{n}\right),$$

where $\beta = c\,(L + L_0)$ for some numerical constant $c > 0$.

2.3.4 Differences and similarities with regression problems

Here we discuss differences and similarities between aggregation of density and regression estimators. Some notation is needed in order to compare these settings.

We first define some notation related to the Density Estimation (DE) framework studied in the present paper. Let X be a random variable with density f absolutely continuous with respect to the measure μ, let $\mathcal{D}^{\mathrm{DE}} = \{f_1, ..., f_M\}$ be a subset of $L^2(\mu)$ and define for all $g \in L^2(\mu)$ and $x \in \mathcal{X}$,

$$\|g\|^2 = \int g^2 d\mu, \quad l_g^{\mathrm{DE}}(x) = \|g\|^2 - 2g(x), \quad g^* = \operatorname*{argmin}_{g \in \mathcal{D}^{\mathrm{DE}}} \|g - f\|^2 = \operatorname*{argmin}_{g \in \mathcal{D}^{\mathrm{DE}}} \mathbb{E}[l_g^{\mathrm{DE}}(X)].$$

Given n i.i.d. observations $X_1, ..., X_n$ and some fixed function g, one can use the empirical risk $P_n(l_g^{\mathrm{DE}}) = \sum_{i=1}^n \frac{1}{n} l_g^{\mathrm{DE}}(X_i)$.

We now define similar notation for the regression problem with the L^2 loss. Let (X, Y) be a random couple valued in $\mathcal{X} \times \mathbf{R}$, let P_X be the probability measure of X, let f be the true regression function defined by $f(x) = \mathbb{E}[Y|X = x]$, let $\mathcal{D}^{\mathrm{R}} = \{f_1, ..., f_M\}$ be a subset of $L^2(P_X)$ and define for all $g \in L^2(P_X)$,

$$\|g\|_{P_X}^2 = \mathbb{E}[g(X)^2], \qquad g^* = \operatorname*{argmin}_{g \in \mathcal{D}^{\mathrm{R}}} \|g - f\|_{P_X}^2.$$

For Regression with Unknown Design (RUD) i.e., when the distribution of the design X is not known to the statistician, a natural choice for the loss function l_g is

$$l_g^{\mathrm{RUD}}(x, y) = (g(x) - y)^2, \qquad \forall x, y \in \mathcal{X} \times \mathbf{R},$$

and the oracle g^* defined above satisfies $g^* = \operatorname{argmin}_{g \in \mathcal{D}^{\mathrm{R}}} \mathbb{E}[l_g^{\mathrm{RUD}}(X, Y)]$. For Regression with Known Design (RKD), the quantity $\|g\|_{P_X}^2$ is accessible for all g. Thus we can define the loss

$$l_g^{\mathrm{RKD}}(x, y) = \|g\|_{P_X}^2 - 2g(x)y, \qquad \forall x, y \in \mathcal{X} \times \mathbf{R},$$

and the oracle g^* satisfies $g^* = \operatorname{argmin}_{g \in \mathcal{D}^{\mathrm{R}}} \mathbb{E}[l_g^{\mathrm{RKD}}(X, Y)]$. Thus, two natural functions l_g arise in the regression context, depending on whether the distribution of the design is known or unknown. Given n i.i.d. observations (X_i, Y_i) with the same distribution as (X, Y), the empirical quantities $P_n(l_g^{\mathrm{RUD}})$ and $P_n(l_g^{\mathrm{RKD}})$ can be used to infer the true regression function f. An estimator constructed using the quantity $P_n(l_g^{\mathrm{RKD}})$ is used, for example, in [98] for the problem of linear and convex aggregation.

Linear or quadratic empirical process. The empirical process $(P_n - P)(l_g - l_{g^*})$ indexed by g plays an important role in the proofs of Theorem 2.2 and Theorem 2.6. This empirical process also appears in the analysis [69] for regression with unknown design with the loss l_g^{RUD}. For density estimation and regression with known design, this empirical process is linear in g:

$$(P_n - P)(l_g^{\mathrm{DE}} - l_{g^*}^{\mathrm{DE}}) = -2(P_n - P)(g - g^*), \qquad (P_n - P)(l_g^{\mathrm{RKD}} - l_{g^*}^{\mathrm{RKD}}) = -2(P_n - P)[(g - g^*)\dot{y}],$$

where the function $\dot{y}(\cdot)$ above is defined by $\forall x, y \in \mathcal{X} \times \mathbf{R}, \dot{y}(x, y) = y$. For regression when the design is unknown, the empirical process is quadratic in the class member g. To control the behavior of this quadratic empirical process, the contraction principle is used in [69], whereas this principle is not needed for density estimation or regression when the distribution of the design is known.

The penalty (2.15) and its coefficient. In the regression problem when the distribution is known, given a dictionary of potential regression functions $\{f_1, ..., f_M\}$, the quantity

$$\sum_{j=1}^{M} \theta_j \|f_j - f_\theta\|_{P_X}^2 , \tag{2.21}$$

is accessible and a procedure similar to the one proposed in Theorem 2.6 and Corollary 2.8 can be constructed, with the penalty coefficient 1/2 which is a natural choice as explained in Remark 2.1. For regression with unknown design, the above penalty cannot be computed: the procedure [69] for the L^2 loss is the estimator $f_{\hat{\theta}}$ where

$$\begin{aligned}\hat{\theta} &= \operatorname*{argmin}_{\theta\in\Lambda^M} \left(P_n\left(l_{f_\theta}^{\mathrm{RUD}}\right) + \nu P_n(f_j - f_\theta)^2 \right), \\ &= \operatorname*{argmin}_{\theta\in\Lambda^M} \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\theta(X_i))^2 + \frac{\nu}{n}\sum_{i=1}^{n}(f_j - f_\theta)^2(X_i) \right),\end{aligned}$$

for some coefficient $\nu \in (0,1)$ and where we chose the uniform prior for clarity. Thus the procedure [69] can be formulated as a penalized procedure where the penalty is the empirical counterpart of (2.21) with the coefficient ν. Although 1/2 is a natural choice for regression with known design and density estimation, for regression with unknown design the expression of the optimal coefficient is more intricate [69, Minimize β in (1.4)].

Sketch of proof for the regression model with known design. In order to show the similarities between density estimation and regression problems when the design is known, we now give the main ideas to derive an oracle inequality similar to Corollary 2.8 for regression with known design. Note that the framework studied in [69] does not cover the estimator defined below, since the function l_g^{RKD} depends on the quantity $\|g\|_{P_X}^2$. Given n i.i.d. observations $(X_1, Y_1), ..., (X_n, Y_n)$, define

$$\hat{\theta} = \operatorname*{argmin}_{\theta\in\Lambda^M} V_n(\theta), \qquad V_n(\theta) = P_n\left(l_{f_\theta}^{\mathrm{RKD}}\right) + \frac{1}{2}\sum_{j=1}^{M}\theta_j \|f_j - f_\theta\|_{P_X}^2 .$$

Analogously to the argument of Proposition 2.7, we note that the function V_n is strongly convex and $V_n(\hat{\theta}) \le V_n(e_{j^*}) - \frac{1}{2}\|f_{j^*} - f_{\hat{\theta}}\|_{P_X}^2$ for any $j^* = 1, ..., M$. As in the proof of Theorem 2.6, this leads to

$$\|f_{\hat{\theta}} - f\|_{P_X}^2 \le \|f_{j^*} - f\|_{P_X}^2 + \max_{k=1,...,M}\delta_k, \qquad \delta_k := (P - P_n)(l_{f_k}^{\mathrm{RKD}} - l_{f_{j^*}}^{\mathrm{RKD}}) - \frac{1}{2}\|f_{j^*} - f_k\|_{P_X}^2 .$$

As explained above, when the distribution of the design is known, the empirical process is linear in $f_k - f_{j^*}$:

$$\delta_k = -2(P - P_n)((f_k - f_{j^*})\dot{y}) - \frac{1}{2}\|f_k - f_{j^*}\|_{P_X}^2 .$$

If for some $b > 0$, $|Y| \le b$ and $\max_{j=1,...,M}|f_j(X)| \le b$ almost surely, then

$$\delta_k \le -2(P - P_n)((f_k - f_{j^*})\dot{y}) - \frac{1}{2b^2}\mathbb{E}[Y^2(f_k(X) - f_{j^*}(X))].$$

Using (2.24) and the union bound on $k = 1, ..., M$, we obtain $\max_{k=1,...,M} \delta_k \leq \beta(x + \log M)/n$ with probability greater than $1 - \exp(-x)$ and thus

$$\|f_{\hat{\theta}} - f\|_{P_X}^2 \leq \|f_{j^*} - f\|_{P_X}^2 + \beta \left(\frac{x + \log M}{n} \right),$$

where $\beta = c\, b^2$ for some numerical constant $c > 0$.

In conclusion, the density estimation framework studied in the present paper is close to the regression problem when the distribution of the design is known, while it presents several differences with the regression problem when the design is not known.

2.4 Minimax optimality in deviation

The goal of this section is to state a minimax optimality result based on the lower bound of Theorem 2.9 and the sharp oracle inequality of Corollary 2.8. In this section, the underlying measure μ is the Lebesgue measure on $\mathbb{R}^d$ for some integer $d \geq 1$.

Minimax optimality in model selection type aggregation is usually defined in expectation [98], by studying the quantity

$$\sup_{\substack{f_j \in \mathcal{F} \\ j=1,...,M}} \inf_{\hat{T}_n} \sup_{f \in \mathcal{F}_d} \left(\mathbb{E} R(\hat{T}_n) - \inf_{j=1,...,M} R(f_j) \right)$$

where the infimum is taken over all estimators $\hat{T}_n$, $\mathcal{F}$ is a class of possible functions for the dictionary and $\mathcal{F}_d$ is the class of all densities satisfying some general constraints.

Let μ be the Lebesgue measure on $\mathbf{R}^d$ and for some $L > 0$, let $\mathcal{F} = \{g \in L^2(\mu), |g|_\infty \leq L\}$ and $\mathcal{F}_d$ be the set of all densities f with respect to μ satisfying $|f|_\infty \leq L$. Then, by an integration argument, Corollary 2.8 and Theorem 2.9 provide the following bounds for some absolute constant $c, C > 0$ and any $M \geq 2, n \geq 1$:

$$c\, \frac{L \log M}{n} \leq \sup_{\substack{f_j \in \mathcal{F} \\ j=1,...,M}} \inf_{\hat{T}_n} \sup_{f \in \mathcal{F}_d} \left(\mathbb{E} R(\hat{T}_n) - \inf_{j=1,...,M} R(f_j) \right) \leq C\, \frac{L \log M}{n}.$$

This shows that $\frac{L \log M}{n}$ is the optimal rate of convergence in expectation for model selection type aggregation under the boundedness assumption.

But our results are stronger that the above optimality in expectation since the deviation term in the sharp oracle inequality of Corollary 2.8 and in the lower bound of Theorem 2.9 are the same up to a numerical constant.

The central quantity when dealing with optimality in deviation is, for $t > 0$,

$$\sup_{\substack{f_j \in \mathcal{F} \\ j=1,...,M}} \inf_{\hat{T}_n} \sup_{f \in \mathcal{F}_d} \mathbb{P} \left(R(\hat{T}_n) - \inf_{j=1,...,M} R(f_j) > t \right).$$

The results of Section 2.3 provide upper and lower bounds for this quantity.

We propose the following definition of minimax optimality in deviation.

Notation 2.1 (Minimax optimality in deviation)**.** Let $\mathcal{F}$ be a subset of $L^2(\mu)$ and $\mathcal{F}_d$ be a set of densities with respect to the measure μ. Let $\mathcal{E}_n$ be a set of estimators. Denote by $\mathbf{P}^{n,M}_{\mathcal{E}_n,\mathcal{F},\mathcal{F}_d}(t)$ the quantity

$$\mathbf{P}^{n,M}_{\mathcal{E}_n,\mathcal{F},\mathcal{F}_d}(t) = \sup_{\substack{f_j\in\mathcal{F}\\ j=1,\dots,M}} \inf_{\hat{T}_n\in\mathcal{E}_n} \sup_{f\in\mathcal{F}_d} \mathbb{P}\left(R(\hat{T}_n) - \inf_{j=1,\dots,M} R(f_j) > t\right).$$

A function $p_{n,M}(\cdot)$ is called optimal tail distribution over $(\mathcal{E}_n,\mathcal{F},\mathcal{F}_d)$ if for any $n\geq 1, M\geq 2$ and any $t>0$,

$$c\, p_{n,M}(c't) \leq \mathbf{P}^{n,M}_{\mathcal{E}_n,\mathcal{F},\mathcal{F}_d}(t) \leq p_{n,M}(t)$$

where $c, c' > 0$ are constants independent of n, M and t.

The following proposition is a direct consequence of Corollary 2.8 and Theorem 2.9.

Proposition 2.10. *Let $M\geq 2, n\geq 1$ and $L>0$. Let $\mathcal{F}=\{g\in L^2(\mathbf{R}^d), |g|_\infty\leq L\}$ and $\mathcal{F}_d$ be the set of all densities f with respect to the Lebesgue measure on $\mathbf{R}^d$ with $|f|_\infty\leq L$. Let $\mathcal{E}_n$ be the set of all estimators. Define*

$$p_{n,M}(t) = M\exp\left(-\frac{3tn}{20L}\right)\mathbf{1}_{[0,L]}(t),$$

where $\mathbf{1}_A$ denotes the indicator function of the set A. Then for all $t>0$,

$$\frac{1}{24}p_{n,M}\left(160\,t\right) \leq \mathbf{P}^{n,M}_{\mathcal{E}_n,\mathcal{F},\mathcal{F}_d}(t) \leq p_{n,M}\left(t\right).$$

Thus, $p_{n,M}(\cdot)$ is an optimal tail distribution over $(\mathcal{E}_n,\mathcal{F},\mathcal{F}_d)$ according to Definition 2.1.

Proof. The regime $t>L$ corresponds to the trivial case where (2.20) holds and $\hat{T}_n^* = 0$ is an optimal estimator. In this regime $p_{n,M}(t)=0$.

For $t\leq L$, by setting $t=\beta\frac{\log(M)+x}{n} = \frac{20L}{3}\frac{\log(M)+x}{n}$ in Corollary 2.8 , we get

$$\mathbf{P}^{n,M}_{\mathcal{E}_n,\mathcal{F},\mathcal{F}_d} \leq p_{n,M}\left(t\right)$$

while Theorem 2.9 implies that

$$\frac{1}{24}p_{n,M}\left(\frac{24\cdot 20}{3}t\right) \leq \mathbf{P}^{n,M}_{\mathcal{E}_n,\mathcal{F},\mathcal{F}_d}(t).$$

□

Similarly, the results of Section 2.2 imply the following proposition.

Proposition 2.11. *Let $M\geq 2, n\geq 1$ and $L>0$. Let $\mathcal{F}=\{g\in L^2(\mathbf{R}^d), |g|_\infty\leq L\}$ and $\mathcal{F}_d$ be the set of all densities f with respect to the Lebesgue measure on $\mathbf{R}^d$ with $|f|_\infty\leq L$. Let $\mathcal{S}_n$ be the set of all selectors, i.e. the measurable functions valued in the discrete set $\{f_1,\dots,f_M\}$. Define*

$$q_{n,M}(t) = M\exp\left(-\frac{t^2 n}{L^2(4\sqrt{2}+8/3)^2}\right)\mathbf{1}_{[0,L]}(t),$$

where $\mathbf{1}_A$ denotes the indicator function of the set A. Then for all $t>0$,

$$\frac{1}{24}q_{n,M}\left(\sqrt{3}(4\sqrt{2}+8/3)\,t\right) \leq \mathbf{P}^{n,M}_{\mathcal{S}_n,\mathcal{F},\mathcal{F}_d}(t) \leq q_{n,M}\left(t\right).$$

Thus, $q_{n,M}(\cdot)$ is an optimal tail distribution over $(\mathcal{S}_n,\mathcal{F},\mathcal{F}_d)$ according to Definition 2.1.

Proof. The regime $t > L$ can be treated similarly as in the proof of Proposition 2.10.

For $t \leq L$, let $t = L(4\sqrt{2} + 8/3)\sqrt{\frac{x+\log M}{n}}$ in Theorem 2.2. For this definition of t and x, $1 \geq \sqrt{\frac{x+\log M}{n}} \geq \frac{x+\log M}{n}$. Then

$$\mathbf{P}^{n,M}_{\mathcal{S}_n,\mathcal{F},\mathcal{F}_d}(t) \leq q_{n,M}(t)$$

and Theorem 2.1 implies

$$\frac{1}{24} q_{n,M}\left(\sqrt{3}(4\sqrt{2} + 8/3)\, t\right) \leq \mathbf{P}^{n,M}_{\mathcal{S}_n,\mathcal{F},\mathcal{F}_d}(t).$$

□

2.5 Proofs

2.5.1 Bias-variance decomposition

As discussed in Section 2.3, the penalty can be viewed as the variance of a random element of the discrete set $\{f_1, \ldots, f_M\}$ and it satisfies the following bias-variance decomposition.

Proposition 2.12. *For any $g \in L^2(\mu)$ and $\theta \in \Lambda^M$,*

$$\sum_{j=1}^{M} \theta_j \|f_j - g\|^2 = \|f_\theta - g\|^2 + \mathrm{pen}(\theta) \tag{2.22}$$

where $\mathrm{pen}(\cdot)$ *is the penalty defined in (2.15).*

Proof. Let η be a random variable with $\mathbb{P}(\eta = j) = \theta_j$ for all $j = 1, \ldots, M$. Denote by $\mathbb{E}_\theta$ the expectation with respect to the random variable η. Then $\mathbb{E}_\theta[f_\eta] = f_\theta$ and the bias-variance decomposition yields

$$\mathbb{E}_\theta \|f_\theta - g\|^2 = \|g - \mathbb{E}_\theta[f_\eta]\|^2 + \mathbb{E}_\theta \|f_\eta - \mathbb{E}_\theta[f_\eta]\|^2,$$

which is exactly the desired result. □

2.5.2 Concentration inequalities

Proposition 2.13. *Let $Y_1, \ldots, Y_n$ be independent random variables, such that almost surely, for all i, $|Y_i - \mathbb{E}Y_i| \leq b$. Then for all $x > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i - \mathbb{E}Y_i > \sqrt{2xv} + \frac{bx}{3}\right) \leq \exp(-x), \tag{2.23}$$

where $v = \sum_{i=1}^{n} \mathbb{V}(Y_i)$.

Proposition 2.13 is close to Bennett and Bernstein inequalities. A proof can be found in [75, Section 2.2.3, (2.20) with $c = b/3$].

The following one-sided concentration inequality is a direct consequence of Proposition 2.13 and the inequality $2\sqrt{uv} \leq \frac{u}{a} + av$ for all $a, u, v > 0$. Under the same assumptions as Proposition 2.13 above, for all $x > 0$ and any $a > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}Y_i - a\mathbb{V}(Y_i) > \left(\frac{1}{2a} + \frac{b}{3}\right)\frac{x}{n}\right) \leq \exp(-x). \tag{2.24}$$

Proposition 2.14. *Let $X_1, \dots, X_n$ be i.i.d. observations drawn from the density f with $|f|_\infty \leq L$. Let $g \in L^2(\mu)$ with $|g|_\infty \leq 4L_0$. Let $\beta = 4L + \frac{8L_0}{3}$. Define*

$$\zeta_n = (P - P_n)g - \frac{1}{8}\|g\|^2 - \frac{\beta}{n}\log\frac{1}{\pi},$$

where the notation P and P_n is defined in (2.9). Then for all $x > 0$,

$$\mathbb{P}\left(\zeta_n > \frac{\beta x}{n}\right) \leq \pi \exp(-x).$$

Proof of Proposition 2.14. As the unknown density f is bounded by L,

$$\mathbb{V}(g(X_1)) \leq P(g^2) = \int g^2 f d\mu \leq L\|g\|^2,$$
$$-\frac{1}{8}\|g\|^2 \leq -\frac{1}{8L}\mathbb{V}(g(X_1)).$$

Thus almost surely

$$\zeta_n \leq (P - P_n)g - \frac{1}{8L}\mathbb{V}(g(X_1)) - \frac{\beta}{n}\log\frac{1}{\pi}.$$

Define n i.i.d. random variables $Y_1, \dots, Y_n$ by

$$Y_i = g(X_i).$$

Almost surely, $|Y_i| \leq 4L_0$ and $|Y_i - \mathbb{E}Y_i| \leq 8L_0$. By applying (2.24) to $Y_1, \dots, Y_n$ with $b = 8L_0$ and $a = \frac{1}{8L}$, we get that for any $t > 0$ with $x = t + \log\frac{1}{\pi}$,

$$\mathbb{P}\left((P - P_n)g - \frac{1}{8L}\mathbb{V}(g(X_1)) > \frac{\beta x}{n}\right) \leq \exp(-x),$$
$$\mathbb{P}\left(\zeta_n > \frac{\beta x}{n}\right) \leq \mathbb{P}\left((P - P_n)g - \frac{1}{8L}\mathbb{V}(g(X_1)) - \frac{\beta}{n}\log\frac{1}{\pi} > \frac{\beta t}{n}\right) \leq \pi\exp(-t).$$

□

2.5.3 Strong convexity

Proof of Proposition 2.7. We will first prove that for any θ, θ',

$$H_n(\theta) - H_n(\theta') = \langle \nabla H_n(\theta'), \theta - \theta' \rangle + \frac{1}{2}\|f_\theta - f_{\theta'}\|^2. \tag{2.25}$$

Using the bias-variance decomposition of (2.22) with $g = 0$, we get

$$\text{pen}(\theta) = \sum_{j=1}^M \theta_j \|f_\theta - f_j\|^2 = -\|f_\theta\|^2 + \sum_{j=1}^M \theta_j \|f_j\|^2.$$

Thus H_n can be rewritten as $H_n(\theta) = \frac{1}{2}\|f_\theta\|^2 + L(\theta)$ where L is affine in θ. If we can prove $N(\theta) - N(\theta') = \langle \nabla N(\theta'), \theta - \theta' \rangle + \|f_\theta - f_{\theta'}\|^2$ where $N(\theta) = \|f_\theta\|^2$, then (2.25) holds. By simple properties of the norm,

$$\|f_\theta\|^2 - \|f_{\theta'}\|^2 = 2\int f_{\theta'}(f_\theta - f_{\theta'})d\mu + \|f_\theta - f_{\theta'}\|^2,$$
$$= 2\theta'^T G(\theta - \theta') + \|f_\theta - f_{\theta'}\|^2,$$

where G is the Gram matrix with elements $G_{j,k} = \int f_j f_k d\mu$ for all $j, k = 1, \dots, M$. The gradient at θ' of the function $\theta \to \|f_\theta\|^2$ is exactly $2G\theta'$ so (2.25) holds.

The function H_n is convex and differentiable. If $\hat{\theta}$ minimizes H_n over the simplex, then for any $\theta \in \Lambda^M$, $\langle \nabla H_n(\hat{\theta}), \theta - \hat{\theta} \rangle \geq 0$ which proves (2.17). □

2.5.4 Tools for lower bounds

Proposition 2.15. *There exists a countable set of functions $\varepsilon_1, \varepsilon_2, \dots$ defined on $[0,1]$ such that for all $j, k > 0$ with $k \neq j$,*

$$\begin{aligned} \forall u \in [0,1), \quad & \varepsilon_j(u) \in \{-1, 1\}, \\ \int_{[0,1]} & \varepsilon_j(x)\varepsilon_k(x)dx = 0, \\ \int_{[0,1]} & \varepsilon_j^2(x)dx = 1. \end{aligned}$$

Furthermore, if U is uniformly distributed on $[0,1]$, then $\varepsilon_1(U), \varepsilon_2(U), \dots$ are i.i.d. Rademacher random variables.

See [54, Definition 3.22] for an explicit construction of these functions and a proof a their properties.

If $P \ll Q$ are two probability measures defined on some measurable space, define their Kullback-Leibler divergence and their χ_2 divergence by

$$K(P,Q) = \int \log\left(\frac{dP}{dQ}\right) dP, \qquad \chi_2(P,Q) = \int \left(\frac{dP}{dQ} - 1\right)^2 dQ.$$

The following comparison holds

$$K(P,Q) \leq \chi_2(P,Q). \tag{2.26}$$

Furthermore, if $n \geq 1$ and $P^{\otimes n}$ denotes the n-product of measures P,

$$K(P^{\otimes n}, Q^{\otimes n}) = nK(P,Q). \tag{2.27}$$

The proofs of (2.26) and (2.27) are given in [99, Lemma 2.7 and page 85].

Lemma 2.16. *Let $(\Omega, \mathcal{A})$ be a measurable space and $m \geq 1$. Let $m \geq 1$ and $A_0, \dots, A_m \in \mathcal{A}$ be disjoint events: $A_j \cap A_k = \emptyset$ for any $j \neq k$. Assume that $Q_0, \dots, Q_m$ are probability measures on $(\Omega, \mathcal{A})$ such that*

$$\frac{1}{m}\sum_{j=1}^{m} K(Q_j, Q_0) \leq \chi < \infty.$$

Then,

$$\max_{j=0,\dots,m} Q_j(\Omega \setminus A_j) \geq \frac{1}{12}\min(1, m\exp(-3\chi)).$$

Lemma 2.16 can be found in [59, Lemma 3]. It is a direct consequence of [99, Proposition 2.3] with $\tau^* = \min(m^{-1}, \exp(-3\chi))$.

Corollary 2.17 (Minimax lower bounds)**.** *Let $n \geq 1$ be an integer and $s > 0$ be a positive number. Let $m \geq 1$ and $q_0, \dots, q_m$ be a family of densities with respect to the same measure μ. Assume that for any $j \neq k$,*

$$\|q_j - q_k\|^2 \geq 4s > 0. \tag{2.28}$$

If $P_k^{\otimes n}$ denotes the product measure associated with n i.i.d. observations drawn from the density q_k, assume that

$$\frac{1}{m}\sum_{j=1}^{m} K(P_j^{\otimes n}, P_0^{\otimes n}) \leq \chi$$

for some finite $\chi > 0$. Then, for any estimator $\hat{T}_n$,

$$\max_{k=0,\dots,m} \mathbb{P}_k^{\otimes n}\left(\|\hat{T}_n - q_k\|^2 \geq s\right) \geq \frac{1}{12}\min(1, m\exp(-3\chi)).$$

Proof of Corollary 2.17. For any estimator $\hat{T}_n$, for any $j = 0, \dots, m$ define the events

$$A_j = \left\{\left\|\hat{T}_n - q_j\right\|^2 < s\right\}.$$

These events are disjoint because of the triangle inequality and (2.28). Applying Lemma 2.16 completes the proof. □

2.5.5 Lower bound theorems

2.5.5.1 Lower bounds with exponential tails

Proof of Theorem 2.9. Let $\varepsilon_2, \dots, \varepsilon_M$ be $M-1$ functions from Proposition 2.15. Consider the dictionary $\{f_1, \dots, f_M\}$ such that for all $(u_1, \dots, u_d) \in \mathbf{R}^d$

$$f_1(u_1, \dots, u_d) = \frac{L}{2}\mathbf{1}_{[0,1]}\left(\frac{L}{2}u_1\right)\prod_{q=2}^{d}\mathbf{1}_{[0,1]}(u_q),$$

and for $j \geq 2$

$$f_j(u_1, \dots, u_d) = \frac{L}{2}\left(1 + \sqrt{\frac{\log(M)+x}{3n}}\varepsilon_j\left(\frac{L}{2}u_1\right)\right)\mathbf{1}_{[0,1]}\left(\frac{L}{2}u_1\right)\prod_{q=2}^{d}\mathbf{1}_{[0,1]}(u_q).$$

Since $\frac{\log M + x}{n} < 3$, these functions are densities and satisfy $|f_j|_\infty < L$.

For any $j \neq k$,

$$\|f_j - f_k\|^2 \geq L\frac{\log(M)+x}{6n} \tag{2.29}$$

and (2.29) is true with equality when $j = 1$. If $P_k^{\otimes n}$ denotes the probability with respect to n i.i.d. random variables with density f_j, the properties (2.26) and (2.27) give that for any $k \geq 2$,

$$\begin{aligned} K(P_k^{\otimes n}, P_1^{\otimes n}) &= nK(P_k^{\otimes 1}, P_1^{\otimes 1}), \\ &\leq n\chi_2(P_k^{\otimes 1}, P_1^{\otimes 1}), \\ &= n\frac{2}{L}\|f_k - f_1\|^2, \\ &= \frac{\log(M)+x}{3}. \end{aligned}$$

Applying Corollary 2.17 with $m = M - 1$ yields that for any estimator $\hat{T}_n$,

$$\sup_{j=1,\dots,M} P_j^{\otimes n}\left(\left\|\hat{T}_n - f_j\right\|^2 \geq L\frac{\log(M)+x}{24n}\right) \geq \frac{1}{12}\min(1, \frac{M-1}{M}\exp(-x)),$$
$$\geq \frac{1}{24}\exp(-x).$$

□

Proof of Theorem 2.1. Let $\varepsilon_1, \dots, \varepsilon_M$ be M functions from Proposition 2.15.

For $(u_1, \dots, u_d) \in \mathbf{R}^d$, we define a dictionary $\{f_1, \dots, f_M\}$ by

$$f_j(u_1, \dots, u_d) = \frac{L}{2}\left(1 + \varepsilon_j\left(\frac{L}{2}u_1\right)\right)\mathbf{1}_{[0,1]}\left(\frac{L}{2}u_1\right)\prod_{q=2}^{d}\mathbf{1}_{[0,1]}(u_q),$$

and we define M densities $\{d_1, \dots, d_M\}$:

$$d_j(u_1, \dots, u_d) = \frac{L}{2}\left(1 + \gamma\varepsilon_j\left(\frac{L}{2}u_1\right)\right)\mathbf{1}_{[0,1]}\left(\frac{L}{2}u_1\right)\prod_{q=2}^{d}\mathbf{1}_{[0,1]}(u_q),$$

for some $\gamma \in (0, \frac{1}{2})$ that will be specified later. Due to the properties of the (ε_j), the following holds for any $j \neq k$

$$\|f_k - d_j\|^2 = \frac{L}{2}(1+\gamma^2),$$
$$\|f_j - d_j\|^2 = \frac{L}{2}(1-\gamma)^2,$$
$$\|d_j - d_k\|^2 = L\gamma^2.$$

Thus if $\hat{S}_n$ is any selector taking values in the discrete set $\{f_1, \dots, f_M\}$:

$$\|\hat{S}_n - d_j\|^2 - \inf_{l=1,\dots,M}\|f_l - d_j\|^2 = \|\hat{S}_n - d_j\|^2 - \|f_j - d_j\|^2 = 2L\gamma\mathbf{1}_{\hat{S}_n \neq f_j}. \quad (2.30)$$

Let $P_k^{\otimes n}$ be the product measure associated with n i.i.d. random variables drawn from the density d_k. Equation (2.30) ensures that with probability $\mathbb{P}_j^{\otimes n}(\hat{S}_n \neq f_j)$, the excess risk is $2L\gamma$.

For any $k \neq 1$, using (2.26) and (2.27), we obtain

$$K(P_k^{\otimes n}, P_1^{\otimes n}) = nK(P_k^{\otimes 1}, P_1^{\otimes 1}),$$
$$\leq n\chi_2(P_k^{\otimes 1}, P_1^{\otimes 1}),$$
$$\leq \frac{4}{L}n\left\|d_k - d_1\right\|^2,$$
$$= 4n\gamma^2,$$

where we used that $d_1(u_1, \dots, u_d) \geq L/4$ almost surely on the common support of d_k and d_1.

Now we choose $\gamma = \frac{1}{2\sqrt{3}}\sqrt{\frac{x+\log M}{n}}$ such that $\forall k \neq 1, K(P_k^{\otimes n}, P_1^{\otimes n}) \leq \frac{x+\log M}{3}$. Let $\hat{S}_n$ be any estimator with values in the discrete set $\{f_1, \dots, f_M\}$. For any $j = 1, \dots, M$, define the event $A_j = \{\hat{S}_n = f_j\}$. The events are disjoint if $f_j \neq f_k$ for

all $j \neq k$ (if this is not satisfied, we can always remove the duplicates). By applying Lemma 2.16 with $m = M-1$ and $\chi = \frac{1}{3}(x + \log M)$, we get

$$\max_{j=1,\dots,M} \mathbb{P}_j^{\otimes n}\left(\hat{S}_n \neq f_j\right) \geq \frac{M-1}{12M}\exp(-x).$$

Since $(M-1)/M \geq 1/2$,

$$\begin{aligned}\max_{j=1,\dots,M} \mathbb{P}_j^{\otimes n}\left(\|\hat{S}_n - d_j\|^2 - \inf_{l=1,\dots,M}\|f_l - d_j\|^2 > 2L\gamma\right) &\geq \frac{M-1}{12M}\exp(-x),\\ &\geq \frac{1}{24}\exp(-x).\end{aligned}$$

□

2.5.5.2 ERM over the convex hull

Proof of Proposition 2.3. By homogeneity, it is enough to prove the case $L = 2$. Let $\phi_1, \dots, \phi_M, \phi_{M+1}$ be $M+1$ functions given by Proposition 2.15. Consider the probability density $f = \mathbf{1}_{[0,1]}$ and the dictionary of $2M+1$ functions

$$\mathcal{D} = \left\{\mathbf{1}_{[0,1]}\right\} \cup \left\{(1 \pm \phi_j\phi_{M+1})\mathbf{1}_{[0,1]}, j = 1, \dots, M\right\}.$$

The true density is in the dictionary thus $\min_{g\in\mathcal{D}}\|f-g\|^2 = 0$. Also, all the elements of the dictionary are uniformly bounded by $L = 2$.

The convex hull of the dictionary is the set

$$\{g_\lambda = (1 + f_\lambda\phi_{M+1})\mathbf{1}_{[0,1]}, \quad \lambda \in \mathbb{R}^M, |\lambda|_1 \leq 1\},$$

where $f_\lambda = \sum_{j=1}^M \lambda_j\phi_j$ and $|\cdot|_1$ is the ℓ_1 norm in $\mathbb{R}^M$.

For all $\lambda \in \mathbb{R}^M$ with $|\lambda|_1 \leq 1$, $\|f - g_\lambda\|^2 = |\lambda|_2^2$ where $|\cdot|_2$ is the ℓ_2 norm in $\mathbb{R}^M$.

Let $\mathcal{L}_\lambda := \|g_\lambda\|^2 - 2g_\lambda + 2f - \|f\|^2 = |\lambda|_2^2 - 2f_\lambda\phi_{M+1}$. Since the empirical process is linear in λ, the proof from [66] can be adapted as follows. Given n i.i.d. observations $X_1, \dots, X_n$ generated by the density f, [66, Lemma 5.4] states that for every $r > 0$, with probability greater than $1 - 6\exp(-C_2M)$,

$$c_0\sqrt{\frac{r}{M}} \leq c_1\sqrt{\frac{rM}{n}} \leq \sup_{\lambda\in\mathbb{R}^M, |\lambda|_2\leq\sqrt{r}} P_n(f_\lambda\phi_{M+1}) \leq c_2\sqrt{\frac{rM}{n}}, \leq c_3\sqrt{\frac{r}{M}},$$

where $c_0, c_1, c_2, c_3 > 0$ are absolute constants.

Let $r \leq 1/M$ that will be specified later (such that if $|\lambda|_2 \leq \sqrt{r}$ then $|\lambda|_1 \leq 1$). On the one hand,

$$\inf_{\lambda\in\mathbb{R}^M, |\lambda|_2\leq\sqrt{r}} P_n\mathcal{L}_\lambda \leq r - 2\sup_{\lambda\in\mathbb{R}^M, |\lambda|_2\leq\sqrt{r}} P_n(f_\lambda\phi_{M+1}).$$

Given that $n \sim M^2$, using the above high probability estimate, there exists a positive absolute constant c_4 such that for all $r \leq c_3^2/(4M)$, with probability greater than $1 - 6\exp(-C_2M)$, $\inf_{\lambda\in\mathbb{R}^M, |\lambda|_2\leq\sqrt{r}} P_n\mathcal{L}_\lambda \leq \sqrt{r}(\sqrt{r} - c_3/\sqrt{M}) \leq -c_4\sqrt{r/M}$, where $c_4 = c_3/2$.

On the other hand, if $\rho \le 1/M$, with probability greater than $1 - 6\exp(-C_2 M)$,

$$\sup_{\lambda\in\mathbb{R}^M, |\lambda|_2\le\sqrt{\rho}} |(P_n - P)\mathcal{L}_\lambda| = 2 \sup_{\lambda\in\mathbb{R}^M, |\lambda|_2\le\sqrt{\rho}} |(P_n - P) f_\lambda \phi_{M+1}| \le 2c_3\sqrt{\frac{\rho}{M}}.$$

Finally, choose r, ρ such that $2c_3\sqrt{\rho/M} < c_4\sqrt{r/M}$ and $\rho > c_5/\sqrt{n}$ for some absolute constant $c_5 > 0$, then with probability greater than $1 - 12\exp(-C_2 M)$,

$$\inf_{\lambda, |\lambda|_2\le\sqrt{\rho}} P_n\mathcal{L}_\lambda \ge - \sup_{\lambda, |\lambda|_2\le\sqrt{\rho}} |(P_n - P)\mathcal{L}_\lambda| \ge -2c_3\sqrt{\frac{\rho}{M}} > -c_4\sqrt{\frac{r}{M}} \ge \inf_{\lambda, |\lambda|_2\le\sqrt{r}} P_n\mathcal{L}_\lambda.$$

Thus with high probability, $\inf_{\lambda,|\lambda|_2\le\sqrt{\rho}} P_n\mathcal{L}_\lambda > \inf_{\lambda,|\lambda|_1\le 1} P_n\mathcal{L}_\lambda$. The inequality is strict so the empirical risk minimizer has a risk greater than ρ. As ρ satisfies $\rho > C_3/\sqrt{n}$, the proof is complete. □

2.5.5.3 Exponential Weights

If $Y_1,...,Y_m$ are i.i.d. with $\mathbb{P}(Y_1 = \pm 1) = 1/2$, then for all $u \in [0, \sqrt{m}/4]$,

$$\frac{1}{15}\exp(-4u^2) \le \mathbb{P}\left(Y_1 + ... + Y_m \ge u\sqrt{m}\right) \le \exp(-u^2/2). \tag{2.31}$$

A proof of the lower bound can be found in [76, 7.3.2] and a standard Chernoff bound provides the upper bound. The following proof uses arguments similar to [32].

Proof of Proposition 2.4. By homogeneity, it is enough to prove the case $L = 1$. Let $\varepsilon_1, \varepsilon_2, \varepsilon_3$ be 3 functions from Proposition 2.15. Let $f = \mathbf{1}_{[0,1]}$ be the unknown density and let

$$f_1 = f + \varepsilon_1, \qquad f_2 = f + (1 + \frac{1}{\sqrt{n}})\varepsilon_2, \qquad f_3 = f_2 + \frac{\alpha}{\sqrt{n}}\varepsilon_3,$$
$$\pi_1 = 1/(8\sqrt{n}), \qquad \pi_2 = 1/(8\sqrt{n}), \qquad \pi_3 = 1 - 1/(4\sqrt{n}),$$

where $0 \le \alpha \le n^{1/4}$ will be specified later. The best function in the dictionary is f_1: $\|f_1 - f\|^2 = \min_{j=1,...,M} \|f_j - f\|^2$.

Let E be the event $\{R_n(f_2) + 2/\sqrt{n} \le R_n(f_1)\}$. By simple algebra,

$$E = \left\{1 + 4\sqrt{n} - 2\sqrt{n}P_n(\varepsilon_2) \le 2n(P_n(\varepsilon_2 - \varepsilon_1)\right\} \supseteq \left\{7\sqrt{n} \le 2n(P_n(\varepsilon_2 - \varepsilon_1)\right\},$$

where for the inclusion we used $1 \le \sqrt{n}$ and $|P_n(\varepsilon_2)| \le 1$. The $2n$ random variables $(\varepsilon_j(X_i))_{j=1,2;\, i=1,...,n}$ are i.i.d. Rademacher random variables, so applying the lower bound of (2.31) with $m = 2n$ and $u = 7\sqrt{2}/4$ yields $\mathbb{P}(E) \ge C_2 > 0$ for some absolute constant C_2. Now set $\alpha^2 = 8\log(2n/C_2)$, and choose N_0 such that for all $n \ge N_0$, $8\log(2n/C_2) > 0$ and $\alpha^2 \le \sqrt{n}$.

Let $F := \{R_n(f_3) \le R_n(f_1)\}$ and define

$$G = \{2(\alpha/\sqrt{n})P_n(\varepsilon_3) \le \alpha^2/n - 2/\sqrt{n}\}.$$

Since $R_n(f_3) = R_n(f_2) + \alpha^2/n - 2(\alpha/\sqrt{n})P_n(\varepsilon_3)$, we have $E \cap G^c \subseteq F$. As $\alpha^2 \le \sqrt{n}$ holds, we have $\alpha^2 - 2\sqrt{n} \le -\alpha^2$ and

$$G \subseteq \{(2(\alpha/\sqrt{n})P_n(\varepsilon_3) \le -\alpha^2/n\} = \{-nP_n(\varepsilon_3) \ge \sqrt{n}\alpha/2\}.$$

The random variable $-nP_n(\varepsilon_j)$ is the sum of n independent Rademacher random variables. Applying the upper bound of (2.31) to $u = \alpha/2$, we have $\mathbb{P}(G) \leq \exp(-\alpha^2/8) = C_2/(2n)$ since $\alpha = 8\log(2n/C_2)$. Now as $F^c \subset E^c \cup G$,

$$\mathbb{P}(E^c \cup F^c) \leq \mathbb{P}(E^c \cup G) \leq (1 - C_2) + \frac{C_2}{2n} \leq 1 - C_2/2 < 1.$$

The probability of the event $E \cap F$ is greater than $C_0 := C_2/2$. On this event, $R_n(f_2) \leq R_n(f_1)$ and $R_n(f_3) \leq R_n(f_1)$ thus

$$\begin{aligned}\hat{\theta}_1^{EW,\beta} &= \frac{\pi_1 \exp(-R_n(f_1)/\beta)}{\pi_1 \exp(-R_n(f_1)/\beta) + \pi_2 \exp(-R_n(f_2)/\beta) + \pi_3 \exp(-R_n(f_3)/\beta)}, \\ &\leq \frac{\pi_1 \exp(-R_n(f_1)/\beta)}{(\pi_1 + \pi_2 + \pi_3)\exp(-R_n(f_1)/\beta)} = \pi_1 = \frac{1}{8\sqrt{n}}.\end{aligned}$$

Let $\theta_1 = \hat{\theta}_1^{EW,\beta}$ for simplicity. As $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ is an orthonormal system,

$$\begin{aligned}\|f_{\hat{\theta}^{EW,\beta}} - f\|^2 - \|f_1 - f\|^2 &\geq \|\theta_1 f_1 + (1-\theta_1)f_2 - f\|^2 - \|f_1 - f\|^2, \\ &= (1-\theta_1)^2 \|f_2 - f\|^2 - (1 - \theta_1^2)\|f_1 - f\|^2, \\ &\geq 2(1-\theta_1)^2/\sqrt{n} + [(1-\theta_1)^2 - (1-\theta_1^2)], \\ &\geq 1/(2\sqrt{n}) - 2\theta_1, \\ &\geq 1/(2\sqrt{n}) - 2/(8\sqrt{n}) \geq 1/(4\sqrt{n}).\end{aligned}$$

□

The proof of Proposition 2.5 is based on estimates from [68] and highlights the similarities between regression with random design and density estimation with the L^2 risk.

Proof of Proposition 2.5. By homogeneity, it is enough to prove the case $L = 1$. The strategy is to construct an example for density estimation such that the calculations from [68, Proof of Theorem A] can be leveraged. Let f_Y be the probability density

$$f_Y(x) = \begin{cases} 1/4 + 1/(2\sqrt{n}) & \text{if } x \in [-2, 0), \\ 1/4 - 1/(2\sqrt{n}) & \text{if } x \in (0, 2], \end{cases}$$

and 0 elsewhere. Let $\{f_1 = \frac{1}{2}\mathbf{1}_{[-2,0)]}, f_2 = \frac{1}{2}\mathbf{1}_{[0,2)}\}$ be the dictionary. Let

$$\mathcal{L}_2(y) := \|f_2\|^2 - 2f_2(y) + 2f_1(y) - \|f_1\|^2, \quad \forall y \in \mathbf{R},$$

and observe that $\mathcal{L}_2(Y) = -X$ where $X = \mathbf{1}_{[0,2)}(Y) - \mathbf{1}_{[-2,0)}(Y)$ so that X satisfies

$$X = \begin{cases} 1 & \text{with probability } 1/2 - 1/\sqrt{n}, \\ -1 & \text{with probability } 1/2 + 1/\sqrt{n}. \end{cases}$$

By definition of $\mathcal{L}_2$,

$$P\mathcal{L}_2 = \mathbb{E}\mathcal{L}_2(Y) = \|f_2 - f_Y\|^2 - \|f_1 - f_Y\|^2.$$

As $P\mathcal{L}_2 = \mathbb{E}[-X] = 2/\sqrt{n} > 0$, f_1 is the best function in the dictionary and $P\mathcal{L}_2$ is the excess risk of f_2. Finally, let

$$\alpha = \frac{\|f_1 - f_2\|^2}{P\mathcal{L}_2} = \frac{\sqrt{n}}{2}.$$

For any $\theta \in [0,1]$, let $f_\theta = \theta f_1 + (1-\theta) f_2$. An explicit calculation of the excess risk of f_θ yields

$$\begin{aligned}\|f_\theta - f_Y\|^2 - \|f_1 - f_Y\|^2 &= \theta^2 \|f_1\|^2 + (1-\theta)^2 \|f_2\|^2 - 2\mathbb{E}[f_\theta(Y)] + 2\mathbb{E}[f_1(Y)] - \|f_1\|^2, \\ &= -\theta(1-\theta)\|f_1 - f_2\|^2 + (1-\theta)\mathbb{E}[-X], \\ &= (1 - \theta - \theta(1-\theta)\alpha) P\mathcal{L}_2.\end{aligned}$$

Given n independent observations $Y_1, ..., Y_n$ with common density f, define $X_i = \mathbf{1}_{[0,2)}(Y_i) - \mathbf{1}_{[-2,0)}(Y_i)$ as above. The exponential weights estimator with temperature β can be written as

$$\hat{f}_\beta^{EW} = \hat{\theta}_1 f_1 + (1 - \hat{\theta}_1) f_2, \qquad \hat{\theta}_1 := \frac{1}{1 + \exp(-(n/\beta)\frac{1}{n}\sum_{i=1}^n [-X_i])},$$

and its excess risk is $\left\|\hat{f}_\beta^{EW} - f_Y\right\|^2 - \|f_1 - f_Y\|^2 = (1 - \hat{\theta}_1 - \hat{\theta}_1(1-\hat{\theta}_1)\alpha) P\mathcal{L}_2$.

The constants α and $P\mathcal{L}_2$, the law of $X_1, ..., X_n, \hat{\theta}_1$ are the same as in [68, Proof of Theorem A], thus the lower bounds in expectation and probability of the quantity $(1 - \hat{\theta}_1 - \hat{\theta}_1(1-\hat{\theta}_1)\alpha)$ in Lecué and Mendelson [68] also hold here and yield the lower bound of Proposition 2.5. □

Chapter 3

Optimal bounds for aggregation of affine estimators

We study the problem of aggregation of estimators when the estimators are not independent of the data used for aggregation and no sample splitting is allowed. If the estimators are deterministic vectors, it is well known that the minimax rate of aggregation is of order $\log(M)$*, where* M *is the number of estimators to aggregate. It is proved that for affine estimators, the minimax rate of aggregation is unchanged: it is possible to handle the linear dependence between the affine estimators and the data used for aggregation at no extra cost. The minimax rate is not impacted either by the variance of the affine estimators, or any other measure of their statistical complexity. The minimax rate is attained with a penalized procedure over the convex hull of the estimators, for a penalty that is inspired from the Q-aggregation procedure. The results follow from the interplay between the penalty, strong convexity and concentration.*

3.1 Introduction

We study the problem of recovering an unknown vector $\boldsymbol{f} = (f_1, ..., f_n)^T \in \mathbb{R}^n$ from noisy observations

$$Y_i = f_i + \xi_i, \qquad i = 1, ..., n, \tag{3.1}$$

where the noise random variables $\xi_1, ..., \xi_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ or i.i.d. subgaussian random variables. We measure the quality of estimation of the unknown vector $\boldsymbol{f}$ with the squared Euclidean norm in $\mathbb{R}^n$:

$$\|\boldsymbol{f} - \hat{\boldsymbol{\mu}}\|_2^2,$$

for any estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{f}$. When the noise random variables are normal, (3.1) is the Gaussian sequence model, which has been extensively studied, see e.g. [57] and the references therein. Several estimators have been proposed to recover the unknown vector $\boldsymbol{f}$ from the observations: the Ordinary Least Squares, the Ridge estimator, the Stein estimator and the procedures based on shrinkage, to name a few. Several of these estimators depend on a parameter that must be chosen carefully to obtain satisfying error bounds. These available estimators have different strengths and weaknesses in different scenarios, so it is important to be able to mimic the best among a given family of estimators, without any assumption on the unknown $\boldsymbol{f}$. The problem of mimicking the best estimator in a given finite set is the problem of model-selection type aggregation, which was introduced in [83, 98]. More

precisely, let $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ be M estimators of $\boldsymbol{f}$ based on the data $\mathbf{y} = (Y_1, ..., Y_n)^T$. The goal is to construct with the same data $\mathbf{y} = (Y_1, ..., Y_n)^T$ a new estimator $\hat{\boldsymbol{\mu}}$ called the aggregate, which satisfies with probability greater than $1 - \delta$ the sharp oracle inequality[1]

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + \text{PRICE}_M(\delta), \tag{3.2}$$

where $\text{PRICE}_M(\cdot)$ is a function of δ that should be small. The term $\text{PRICE}_M(\cdot)$ will be referred to as the price to pay for aggregating the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$. If the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ are deterministic vectors, the price to pay for aggregating these estimators is of order $\sigma^2 \log(M/\delta)$ and (3.2) is satisfied for an estimator $\hat{\boldsymbol{\mu}}$ based on Q-aggregation [32]. Considering deterministic estimators is of interest if two independent samples are available, so that $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ are based on the first sample while aggregation is performed using the second sample. Then the first sample can be considered as frozen at the aggregation step (for more details see [96]). If the estimators are random (dependent on the data $\mathbf{y}$ used for aggregation), two natural questions arise.

1. Does the price to pay for aggregation increase because of the dependence between $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ and the data $\mathbf{y}$, or is it still of order $\sigma^2 \log(M/\delta)$? Is there an extra price to pay to handle the dependence?

2. A natural quantity that captures the statistical complexity of a given estimator $\hat{\boldsymbol{\mu}}_j$ is the variance defined by $\mathbb{E}|\hat{\boldsymbol{\mu}}_j - \mathbb{E}\hat{\boldsymbol{\mu}}_j|_2^2$. When the estimators are deterministic, their variances are all zero. Now that the estimators are random, does the price to pay for aggregation depend on the statistical complexities of the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$, for example through their variances? Is it harder to aggregate estimators with large statistical complexities?

The goal of this paper is to answer these questions for affine estimators.

Among the procedures available to estimate $\boldsymbol{f}$, several are linear in the observations $Y_1, ..., Y_n$. It is the case for the Least Squares and the Ridge estimators, whereas the shrinkage estimators and the Stein estimator are non-linear functions of the observations. Examples of estimators that are linear or affine in the observations is given in [35, Section 1.2], [2] and references therein. An affine estimator is of the form $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j$ for a deterministic matrix A_j of size $n \times n$ and a deterministic vector $\boldsymbol{b}_j \in \mathbb{R}^n$. The linearity of the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ makes it possible to explicitly treat the dependence between the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ and the data $\mathbf{y} = (Y_1, ..., Y_n)^T$ used to aggregate them. Donoho et al. [40] proved that for orthosymmetric quadratically convex sets (which include all ellipsoids and hyperrectangles), the minimax risk among all linear estimators is within 25% of the minimax risk among all estimators.

The papers [70, 35, 33] derived different procedures that satisfy sharp oracle inequalities for the problem of aggregation of affine estimators when the noise random variables are Gaussian. Leung and Barron [70], Dalalyan and Salmon [35] proposed an estimator $\hat{\boldsymbol{\mu}}^{EW}$ based on exponential weights, for which the following sharp oracle inequality holds in expectation:

$$\mathbb{E}|\boldsymbol{f} - \hat{\boldsymbol{\mu}}^{EW}|_2^2 \leq \min_{j=1,...,M} \mathbb{E}|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 8\sigma^2 \log M,$$

[1] By sharp, we mean that the constant in front of the term $\min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2$ is 1.

under the assumption that all A_j are orthoprojectors (orthogonal projection matrices, cf (3.4)), or under a strong commutativity assumption on the matrices A_j. The constant 8 can be reduced to 4 if all A_j are orthoprojectors. If the matrices A_j are not symmetric, [35] achieved a similar oracle inequality by symmetrizing the affine estimators before the aggregation step, which suggests that the symmetry assumption can be relaxed. Although the estimator $\hat{\boldsymbol{\mu}}^{EW}$ achieves this inequality in expectation, it was shown in [3, 32] that it cannot achieve a similar result in deviation, with an unavoidable error term of order $\sqrt{n}$. In Dai et al. [33], a sharp oracle inequality in deviation is derived for an estimator $\hat{\boldsymbol{\mu}}^{Q}$ based on Q-aggregation [90, 32]. Namely, [33] proves that if the matrices $A_1, ..., A_M$ are symmetric and positive semi-definite, the estimator $\hat{\boldsymbol{\mu}}^{Q}$ satisfies with probability greater than $1-\delta$:

$$|\boldsymbol{f} - \hat{\boldsymbol{\mu}}^{Q}|_2^2 \leq \min_{j=1,...,M} \Big(|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 4\sigma^2 \mathrm{Tr}(A_j) \Big) + C\sigma^2 \log(M/\delta), \tag{3.3}$$

where the constant C is proportional to the largest operator norm of the matrices $A_1, ..., A_M$. The term $4\sigma^2\mathrm{Tr}(A_j)$ is intimately linked to the statistical complexity of the estimator $\hat{\boldsymbol{\mu}}_j = A_j\mathbf{y} + \boldsymbol{b}_j$. For instance, the variance of $\hat{\boldsymbol{\mu}}_j$ is $\mathbb{E}|\hat{\boldsymbol{\mu}}_j - \mathbb{E}\hat{\boldsymbol{\mu}}_j|_2^2 = \sigma^2\mathrm{Tr}(A_j^T A_j)$. If $\hat{\boldsymbol{\mu}}_j$ is a Least Squares estimator, A_j is an orthoprojector, and the variance becomes $\sigma^2 \mathrm{Tr} A_j$. Thus, the statistical complexity of the estimator $\hat{\boldsymbol{\mu}}_j$ clearly appears in the remainder term of the oracle inequality (3.3) proved in [33]. Thus, one may think that the price to pay for aggregating affine estimators, i.e. the function $\textsc{price}_M(\delta)$ in (3.2), depends on the statistical complexity of the estimators to aggregate.

The bound (3.3) may lead to the conclusion that the price to pay for aggregation of affine estimators can be substantially larger than $\sigma^2 \log(M/\delta)$ which is the price for aggregating deterministic vectors. Indeed, the extra term $4\sigma^2\mathrm{Tr}(A_j)$ may be large in common situation where the trace of some matrices A_j is large. For example, if one aggregates the estimators $\hat{\boldsymbol{\mu}}_1 = \lambda_1\mathbf{y}, ..., \hat{\boldsymbol{\mu}}_M = \lambda_M\mathbf{y}$, for some positive real numbers $\lambda_1, ..., \lambda_M$, then the remainder term $4\sigma^2\mathrm{Tr}(A_j)$ in the above oracle inequality is of order $\sigma^2 n\lambda_j$ for each $j = 1, ..., M$, which can be greater than the optimal rate $\sigma^2 \log M$. This term $4\sigma^2\mathrm{Tr}(A_j)$ makes the oracle inequality (3.3) suitable only for scenarios where the matrices A_j have small trace. But more importantly, the term $\sigma^2 \mathrm{Tr} A_j$ suggests that the price to pay for aggregating affine estimators increases with the statistical complexities of the estimators to aggregate.

The results discussed above rely on specific assumptions on the matrices $A_1, ..., A_M$ [70, 35, 33]. This raises a third question, although not as important as the two questions above:

3. Does the nature of the matrices $A_1, ..., A_M$ have an impact on the price to pay to aggregate these affine estimators? Is the price in (3.2) substantially smaller if the matrices are orthoprojectors, semi-positive definite or symmetric?

The main contribution of the present paper is to answer the three questions raised above:

1. It is proved in Theorem 3.1 that a penalized procedure over the simplex satisfies the sharp oracle inequality (3.2) with $\textsc{price}_M(\delta) = c\sigma^2 \log(M/\delta)$ for some absolute constant $c > 0$. This price is of the same order as for the problem of aggregation of deterministic vectors. Thus the dependence between the estimators and the data used to aggregate them induces no extra cost.

2. The form of the affine estimators to aggregate has no impact on the price to pay for aggregation. In particular, the sharp oracle inequalities of the present paper do not involve quantities dependent on A_j such as $\sigma^2 \mathrm{Tr} A_j$.

3. The only assumption made on the matrices $A_1, ..., A_M$ is that $|||A_j|||_2 \leq 1$ for all $j = 1, ..., M$, where $|||\cdot|||_2$ is the operator norm. All other assumptions on the matrices $A_1, ..., A_M$ can be dropped, in particular the matrices can be non-symmetric and have negative eigenvalues.

The paper is organized as follows. In Section 3.1.1 we define the notation used throughout the paper. Section 3.2 defines a penalized procedure over the simplex and shows that it achieves sharp oracle inequalities in deviation for aggregation of affine estimators. The role of the penalty is studied in Section 3.3 and Section 3.4. Prior weights are considered in Section 3.5. Section 3.6 shows that the estimator is robust to variance misspecification and to non-Gaussianity of the noise. Some examples are given in Section 3.7. Section 3.8 is devoted to the proofs.

3.1.1 Notation

Let $\boldsymbol{f} = (f_1, ..., f_n)^T \in \mathbb{R}^n$ be an unknown regression vector. We observe n random variables (3.1) where $\xi_1, ..., \xi_n$ are subgaussian random variables, with $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i^2] = \sigma^2$. It can be rewritten in the vector form $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\xi}$ where $\mathbf{y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{f} = (f_1, ..., f_n)^T$ and $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$.

For any estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{f}$, we measure the quality of estimation of $\boldsymbol{f}$ with the loss $|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2$, where $|\cdot|_2$ is the Euclidean norm in $\mathbb{R}^n$. Let $M \geq 2$. We consider M affine estimators of the form

$$\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j, \qquad j = 1, ..., M.$$

The matrices $A_1, ..., A_M$ and the vectors $\boldsymbol{b}_1, ..., \boldsymbol{b}_M \in \mathbb{R}^n$ are deterministic. Define the simplex in $\mathbb{R}^M$:

$$\Lambda^M = \left\{ \boldsymbol{\theta} \in \mathbb{R}^M, \quad \sum_{j=1}^M \theta_j = 1, \quad \forall j = 1 \dots M, \ \ \theta_j \geq 0 \right\}.$$

For any $\boldsymbol{\theta} \in \Lambda^M$, let

$$A_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j A_j, \qquad \boldsymbol{b}_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j \boldsymbol{b}_j, \qquad \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = A_{\boldsymbol{\theta}} \mathbf{y} + \boldsymbol{b}_{\boldsymbol{\theta}}.$$

Let $\boldsymbol{e}_1, ..., \boldsymbol{e}_M$ be the vectors of the canonical basis in $\mathbb{R}^M$. Then $\hat{\boldsymbol{\mu}}_j = \hat{\boldsymbol{\mu}}_{\boldsymbol{e}_j}$ for all $j = 1, ..., M$.

An orthoprojector is an $n \times n$ matrix P such that

$$P = P^T = P^2. \tag{3.4}$$

Denote by $I_{n \times n}$ the $n \times n$-identity matrix. For any $n \times n$ real matrix $A = (a_{i,j})_{i,j=1,...,n}$, define the operator norm of A, the Frobenius (or Hilbert-Schmidt) norm of A and the nuclear norm of A respectively by:

$$|||A|||_2 = \sup_{x \neq 0} \frac{|Ax|_2}{|x|_2}, \quad \|A\|_{\mathrm{F}} = \sqrt{\sum_{i,j=1,...,n} a_{i,j}^2}, \quad \|A\|_1 = \mathrm{Tr}\left(\sqrt{A^T A}\right).$$

The following inequalities hold for any two squared matrices M, M':

$$|||MM'|||_2 \leq |||M|||_2 |||M'|||_2, \qquad \|MM'\|_{\mathrm{F}} \leq |||M|||_2 \, \|M'\|_{\mathrm{F}} \,. \tag{3.5}$$

Finally, denote by log the natural logarithm with $\log(e) = 1$.

3.2 A penalized procedure on the simplex

For any $\boldsymbol{\theta} \in \Lambda^M$ define

$$C_p(\boldsymbol{\theta}) := |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 - 2\mathbf{y}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + 2\sigma^2 \mathrm{Tr}(A_{\boldsymbol{\theta}}), \tag{3.6}$$

which is Mallows [73] C_p-criterion. Next, define

$$H_{\mathrm{pen}}(\boldsymbol{\theta}) = C_p(\boldsymbol{\theta}) + \tfrac{1}{2}\mathrm{pen}(\boldsymbol{\theta}), \tag{3.7}$$

where

$$\mathrm{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{M} \theta_j |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_j|_2^2. \tag{3.8}$$

We consider the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}}$ where

$$\hat{\boldsymbol{\theta}}_{\mathrm{pen}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Lambda^M} H_{\mathrm{pen}}(\boldsymbol{\theta}). \tag{3.9}$$

The function H_{pen} is quadratic and convex (cf. Lemma 3.14). Minimizing H_{pen} over the simplex is a convex quadratic program for which efficient algorithms are available. The convexity of H_{pen} also proves that $\hat{\boldsymbol{\theta}}_{\mathrm{pen}}$ is well defined, although it may not be unique (for example if all $\hat{\boldsymbol{\mu}}_j$ are the same then H_{pen} is constant on the simplex).

We now explain the meaning of the terms that appear in (3.7). If $\boldsymbol{\theta}$ is fixed, $C_p(\boldsymbol{\theta})$ is an unbiased estimate of the quantity

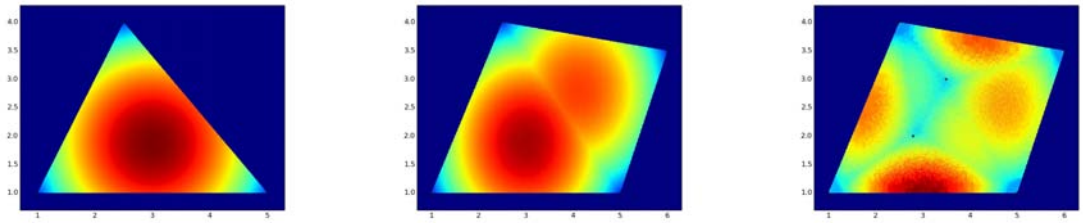
$$R(\boldsymbol{\theta}) := |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 - 2\boldsymbol{f}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 - |\boldsymbol{f}|_2^2, \tag{3.10}$$

which is the quantity of interest $|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2$ up to the additive constant $|\boldsymbol{f}|_2^2$.

The penalty (3.8) is borrowed from the Q-aggregation procedure, which is a powerful tool to derive sharp oracle inequalities in deviation when the loss is strongly convex [90, 32, 69, 12]. Since the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ depend on the data, the penalty (3.8) is data-driven, which is not the case if $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ are deterministic vectors as in [32]. In order to give some geometric insights on the penalty (3.8), let $c \in \mathbb{R}^n$ be a solution of M linear equations $2c^T \hat{\boldsymbol{\mu}}_j = |\hat{\boldsymbol{\mu}}_j|_2^2, j = 1, ..., M$, and assume only in the rest of this paragraph that such a solution exists, even though this assumption cannot be fulfilled for $M > n$. Then

$$\mathrm{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{M} \theta_j |\hat{\boldsymbol{\mu}}_j|_2^2 - |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 = 2c^T \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 = |c|_2^2 - |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - c|_2^2. \tag{3.11}$$

We can write $\mathrm{pen}(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})$ for some function g defined on the convex hull of $\{\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M\}$. Equation (3.11) shows that the level sets of the function g are Euclidean balls centered at c. The function g is non-negative, it is minimal at the extreme points $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ since $g(\hat{\boldsymbol{\mu}}_j) = 0$ for all $j = 1, ..., M$ and g is maximal at the projection of c on the convex hull of $\{\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M\}$. Intuitively, the penalty (3.8) pushes $\boldsymbol{\theta}$ away from the center of the simplex towards the vertices. Thus, the level sets of the function $\boldsymbol{\theta} \to \mathrm{pen}(\boldsymbol{\theta})$ in $\mathbb{R}^M$ are ellipsoids centered at $\boldsymbol{\theta}_c$, where $\boldsymbol{\theta}_c$ is the unique point in $\mathbb{R}^M$ such that $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_c} = c$. If $M > n$ or if the vector c is not well defined, the level sets of $\mathrm{pen}(\cdot)$ are more intricate and cannot be described in such a simple way.



(a) 3 points in $\mathbf{R}^2$. (b) 4 points in $\mathbf{R}^2$. (c) 6 points in $\mathbf{R}^2$.

Figure 3.1: Penalty (3.8) heatmaps. Largest penalty in red, smallest in blue.

Theorem 3.1 (Main result). *Let $M \geq 2$. For $j = 1, ..., M$, consider the affine estimators $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j$ and assume that $|||A_j|||_2 \leq 1$. Assume that the noise random variables $\xi_1, ...\xi_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Let $\hat{\boldsymbol{\theta}}_{\text{pen}}$ be the estimator defined in (3.9). Then for all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$ satisfies with probability greater than $1 - \exp(-x)$,*

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 30\sigma^2(x + 2\log M). \tag{3.12}$$

Furthermore,

$$\mathbb{E}\left[|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2\right] \leq \mathbb{E}\left[\min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2\right] + 60\sigma^2 \log(M). \tag{3.13}$$

The sharp oracle inequality in deviation given in [33] presents an additive term proportional to $\sigma^2 \text{Tr}(A_j)$, as in (3.3). An improvement of the present paper is the absence of this additive term which can be large for matrices A_j with large trace. Our analysis shows that the quantities $\sigma^2 \text{Tr}(A_j)$ are not meaningful for the problem of aggregation of affine estimators, and Theorem 3.1 improves upon the earlier result of [33].

We relax all assumptions on the matrices $A_1, ..., A_M$, for instance they may be non-symmetric and have negative eigenvalues. The above result shows that the restrictions on the matrices $A_1, ..., A_M$ introduced in [70, 35, 33] are not intrinsic to the problem of aggregation of affine estimators.

An estimator of the form $B_j \mathbf{y} + \boldsymbol{b}_j$ with $|||B_j|||_2 > 1$ and $\boldsymbol{b}_j \in \mathbb{R}^n$ is inadmissible in the sense that there exists a matrix $A_j = A_j(B_j)$ such that

$$|||A_j|||_2 \leq 1, \qquad \mathbb{E}\left[|A_j \mathbf{y} + \boldsymbol{b}_j - \boldsymbol{f}|_2^2\right] \leq \mathbb{E}\left[|B_j \mathbf{y} + \boldsymbol{b}_j - \boldsymbol{f}|_2^2\right] \tag{3.14}$$

for all $\boldsymbol{f} \in \mathbb{R}^n$, cf. Cohen [31]. Let $B_1, ..., B_M$ be real matrices and $\boldsymbol{b}_1, ..., \boldsymbol{b}_M \in \mathbb{R}^n$ be deterministic vectors. We now define the matrices $A_j = A_j(B_j)$ in the following way. If $|||B_j|||_2 > 1$ then A_j is a matrix such that (3.14) holds and if $|||B_j|||_2 \leq 1$, set $A_j = B_j$. By Theorem 3.1, the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$ that aggregates the improved estimators $(A_j \mathbf{y} + \boldsymbol{b}_j)_{j=1,...,M}$ satisfies

$$\begin{aligned}
\mathbb{E}|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 &\leq \min_{j=1,...,M} \mathbb{E}\left[|A_j \mathbf{y} + \boldsymbol{b}_j - \boldsymbol{f}|_2^2\right] + 60\sigma^2 \log(M), \\
&\leq \min_{j=1,...,M} \mathbb{E}\left[|B_j \mathbf{y} + \boldsymbol{b}_j - \boldsymbol{f}|_2^2\right] + 60\sigma^2 \log(M).
\end{aligned}$$

Thus, we obtain a sharp oracle inequality in expectation without the assumption $\max_{j=1,...,M} |||B_j|||_2 \leq 1$ if the estimators $(B_j\mathbf{y} + \boldsymbol{b}_j)_{j=1,...,M}$ are pre-improved by transformation to $(A_j\mathbf{y} + \boldsymbol{b}_j)_{j=1,...,M}$ where $|||A_j|||_2 \leq 1$.

The next proposition shows that the bounds of Theorem 3.1 are optimal in a minimax sense. For any $\boldsymbol{f} \in \mathbb{R}^n$ we denote by $\mathbb{P}_{\boldsymbol{f}}$ the probability measure of the random variable $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\xi}$. A lower bound for aggregation of deterministic vectors was proved in [92, Theorem 5.4 with $S = 1$]. This lower bound implies the following result.

Proposition 3.2. *There exist absolute constants $c^*, C^*, p^* > 0$ such that the following holds. For all $M, n \geq C^*$, there exist $\boldsymbol{b}_1, ..., \boldsymbol{b}_M \in \mathbb{R}^n$ and orthoprojectors $A_1, ..., A_M$ of rank one such that*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{f} \in \mathbb{R}^n} \mathbb{P}_{\boldsymbol{f}} \left(|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 - \min_{k=1,...,M} |\boldsymbol{b}_k - \boldsymbol{f}|_2^2 \geq c^* \sigma^2 \log(M) \right) \geq p^*, \tag{3.15}$$

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{f} \in \mathbb{R}^n} \mathbb{P}_{\boldsymbol{f}} \left(|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 - \min_{k=1,...,M} |A_k\mathbf{y} - \boldsymbol{f}|_2^2 \geq c^* \sigma^2 \log(M) \right) \geq p^*, \tag{3.16}$$

where the infima are taken over all estimators $\hat{\boldsymbol{\mu}}$.

This implies that the bounds of Theorem 3.1 are rate minimax in terms of the aggregation price. The lower bound can be constructed either with a dictionary of deterministic vectors (cf. (3.15)), or with a dictionary of orthoprojectors of rank one (cf. (3.16)).

3.3 The penalty (3.8) improves upon model selection based on C_p

In order to explain the role of the penalty (3.8) for the problem of aggregation of affine estimators, consider first the standard empirical risk minimization scheme based on the C_p criterion. Define $\hat{J}$ as

$$\hat{J} \in \operatorname*{argmin}_{j=1,...,M} C_p(\boldsymbol{e}_j), \tag{3.17}$$

where $C_p(\cdot)$ is defined in (3.6). Using that $C_p(\boldsymbol{e}_{\hat{J}}) \leq C_p(\boldsymbol{e}_k)$ for all $k = 1, ..., M$ together with the definition of $C_p(\cdot)$ and $R(\cdot)$ given in (3.6) and (3.10), the following holds almost surely:

$$|\hat{\boldsymbol{\mu}}_{\hat{J}} - \boldsymbol{f}|_2^2 \leq \min_{k=1,...,M} |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \max_{j,k=1,...,M} \Delta_{jk}, \tag{3.18}$$

where $\Delta_{jk} := C_p(\boldsymbol{e}_k) - C_p(\boldsymbol{e}_j) - (R(\boldsymbol{e}_k) - R(\boldsymbol{e}_j))$. Thus, it is possible to prove an oracle inequality for the estimator $\hat{\boldsymbol{\mu}}_{\hat{J}}$ if we can control the quantities Δ_{jk} uniformly over all pairs $j, k = 1, ..., M$. These quantities can be rewritten as

$$\Delta_{jk} = 2\boldsymbol{\xi}^T((A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k) + 2\left(\boldsymbol{\xi}^T(A_j - A_k)\boldsymbol{\xi} - \sigma^2 \mathrm{Tr}(A_j - A_k)\right). \tag{3.19}$$

Two stochastic terms appear in Δ_{jk}. The first is a centered Gaussian random variable with variance $4\sigma^2|(A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k|_2^2$. The second is a centered quadratic form in $\boldsymbol{\xi}$, and it can be shown that its variance is of order $\sigma^4 \|A_j - A_k\|_{\mathrm{F}}^2$. This quadratic term is sometimes called a Gaussian chaos of order 2. The deviations of these

two terms are governed by the following concentration inequalities. For any vector $\boldsymbol{v} \in \mathbb{R}^n$, a standard Gaussian tail bound gives

$$\mathbb{P}\left(\boldsymbol{v}^T\boldsymbol{\xi} > \sigma|\boldsymbol{v}|_2\sqrt{2x}\right) \leq \exp(-x), \qquad \forall x > 0. \tag{3.20}$$

For the Gaussian chaos of order 2, the following is proved in [20, Example 2.12].

Lemma 3.3. *Assume that $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$. For any squared matrix B of size n,*

$$\mathbb{P}\left(\boldsymbol{\xi}^T B\boldsymbol{\xi} - \sigma^2 \mathrm{Tr}B > 2\sigma^2 \left\|B\right\|_{\mathrm{F}} \sqrt{x} + 2\sigma^2 |||B|||_2 x\right) \leq \exp(-x), \tag{3.21}$$

where $\sigma^2\mathrm{Tr}B = \mathbb{E}[\boldsymbol{\xi}^T B\boldsymbol{\xi}]$.

We set $\boldsymbol{v} = 2((A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k)$ and $B = 2(A_k - A_j)$ to study the deviations of the random variable Δ_{jk}. If $|||A_j - A_k|||_2$ is small, (3.20) and (3.21) yield that the deviations of Δ_{jk} are of order of the two quantities

$$\sigma|(A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k|_2, \qquad \sigma^2 \left\|A_j - A_k\right\|_{\mathrm{F}}, \tag{3.22}$$

i.e., the standard deviations of the two terms in Δ_{jk}. The concentration inequalities (3.20) and (3.21) are known to be tight [62], thus there is little hope to bound the deviations of Δ_{jk} independently of $\boldsymbol{f}$, A_j and A_k in order to prove a sharp oracle inequality. It is possible to refine the above analysis and to prove the following oracle inequality, though with a leading constant greater than 1.

Proposition 3.4. *There exist absolute constants $c, C > 0$ such that the following holds. Assume that $|||A_j|||_2 \leq 1$ for all $j = 1, ..., M$. Let $0 < \varepsilon < c$ and let $\hat{J}$ be the estimator defined in (3.17). For all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}_{\hat{J}}$ satisfies with probability greater than $1 - 2\exp(-x)$*

$$|\hat{\boldsymbol{\mu}}_{\hat{J}} - \boldsymbol{f}|_2^2 \leq (1+\varepsilon) \min_{k=1,...,M} |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + C\sigma^2(x + 2\log M)/\varepsilon.$$

The proof of Proposition 3.4 is given in the supplementary material. The estimator $\hat{\boldsymbol{\mu}}_{\hat{J}}$ fails to achieve a sharp oracle inequality with a remainder term of order $\sigma^2 \log M$, and this drawback cannot be repaired for all procedures of the form $\hat{\boldsymbol{\mu}}_{\hat{K}}$ where $\hat{K}$ is an estimator valued in $\{1, ..., M\}$. Indeed, it is proved in [46, Section 6.4.2 and Proposition 6.1] that there exist $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^n$ and orthoprojectors A_1, A_2 such that for any estimator $\hat{K}$ valued in $\{1, 2\}$,

$$\sup_{\boldsymbol{f}\in\{\boldsymbol{f}_1,\boldsymbol{f}_2\}} \left(\mathbb{E}|A_{\hat{K}}\mathbf{y} - \boldsymbol{f}|_2^2 - \min_{j=1,2} \mathbb{E}|A_j\mathbf{y} - \boldsymbol{f}|_2^2\right) \geq \sigma^2\sqrt{n}/4, \tag{3.23}$$

provided that n is larger than some absolute constant. Inspection of the proof of this result reveals that

$$\sigma|(A_2 - A_1)\boldsymbol{f} + \boldsymbol{b}_2 - \boldsymbol{b}_1|_2 \geq \sigma^2\sqrt{n}, \qquad \forall \boldsymbol{f} \in \{\boldsymbol{f}_1, \boldsymbol{f}_2\},$$

where we set $\boldsymbol{b}_1 = \boldsymbol{b}_2 = 0$. Thus, this lower bound of order $\sqrt{n}$ is related to the Gaussian component of the random variable Δ_{12}, i.e., to the term $\boldsymbol{\xi}^T((A_1 - A_2)\boldsymbol{f} + \boldsymbol{b}_1 - \boldsymbol{b}_2)$, cf. (3.19).

The procedure $\hat{\boldsymbol{\mu}}_{\hat{J}}$ fails to achieve a sharp oracle inequality because the variances of the two components of Δ_{jk} may be large and cannot be controlled. The role of the penalty (3.8) is exactly to control the deviations of Δ_{jk} by controlling the terms (3.22). The following proposition makes this precise.

Proposition 3.5. *Let $\hat{\boldsymbol{\theta}}_{\text{pen}}$ be the estimator (3.9). Then almost surely,*

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 \leq \min_{q=1,\dots,M} \left(|\hat{\boldsymbol{\mu}}_q - \boldsymbol{f}|_2^2\right) + \max_{j,k=1,\dots,M} \left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right), \tag{3.24}$$

where Δ_{jk} is the quantity (3.19). Furthermore, for all $j, k = 1, \dots, M$,

$$\mathbb{E}\left[\frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right] = \frac{1}{2}|(A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k|_2^2 + \frac{\sigma^2}{2}\|A_j - A_k\|_{\text{F}}^2. \tag{3.25}$$

The proof of (3.24) is given in Section 3.4 below. A bias-variance decomposition directly yields (3.25), since $\mathbb{E}[\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k] = (A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k$ and $\mathbb{E}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k - \mathbb{E}[\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k]|_2^2 = \mathbb{E}|(A_j - A_k)\boldsymbol{\xi}|_2^2 = \sigma^2\|A_j - A_k\|_{\text{F}}^2$.

Compared with (3.18), the right hand side of (3.24) presents the quantities $-\frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2$. We will explain below that these quantities appear because of the interplay between the penalty (3.8) and the strong convexity of H_{pen}.

From (3.24), an outline of the proof of Theorem 3.1 is as follows. By combining the simple inequality (3.54) and Proposition 3.12 below, we will prove that for any pair (j, k) we have

$$\mathbb{E}\exp\left(\lambda_0\left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right)\right) \leq 1$$

for $\lambda_0 = (30\sigma^2)^{-1}$ if the noise $\boldsymbol{\xi}$ has distribution $\mathcal{N}(0, \sigma^2 I_{n\times n})$. Thus, one has

$$\mathbb{E}\exp\left(\lambda_0 \max_{j,k=1,\dots,M}\left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right)\right) \leq M^2.$$

Then, Jensen's inequality yields (3.13) while a Chernoff bound yields (3.12). This explains the success of the penalty (3.8) for the problem of model selection type aggregation: the penalty and the strong convexity of H_{pen} provide the quantity $-\frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2$, and this quantity is exactly what is needed to control the deviations of the random variable Δ_{jk}.

3.4 Strong convexity and the penalty (3.8)

To further understand the interplay between the penalty (3.8) and the strong convexity of H_{pen}, we now give the proof of (3.24).

Proof of (3.24). Let $k = 1, \dots, M$ be fixed. The simplex Λ^M is a convex set and the function H_{pen} is convex, hence we have

$$\nabla H_{\text{pen}}(\hat{\boldsymbol{\theta}}_{\text{pen}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}_{\text{pen}}) \geq 0,$$

cf. [21, Section 4.2.3, equation (4.21)]. Inequality (3.24) follows from

$$\begin{aligned}
& |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 \\
\leq\ & |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \nabla H_{\text{pen}}(\hat{\boldsymbol{\theta}}_{\text{pen}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}_{\text{pen}}), && (3.26)\\
=\ & \sum_{j=1}^M \hat{\theta}_{\text{pen},j}\left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right), && (3.27)\\
\leq\ & \max_{j=1,\dots,M}\left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right). && (3.28)
\end{aligned}$$

Equality (3.27) is obtained by simple algebra while (3.28) is a consequence of $\sum_{j=1}^M \hat{\theta}_{\text{pen},j} = 1$ and $\hat{\theta}_{\text{pen},j} \geq 0$ for all $j = 1, \dots, M$. □

It is possible to interpret this argument in light of the interplay between strong convexity and the penalty (3.8). The right hand side of (3.26) satisfies

$$\begin{aligned} & |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \nabla H_{\mathrm{pen}}(\hat{\boldsymbol{\theta}}_{\mathrm{pen}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}_{\mathrm{pen}}) \\ = & \sum_{j=1}^M \hat{\theta}_{\mathrm{pen},\,j}\Delta_{jk} - \frac{1}{2}\left[\mathrm{pen}(\hat{\boldsymbol{\theta}}_{\mathrm{pen}}) + |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \hat{\boldsymbol{\mu}}_k|_2^2\right]. \end{aligned}$$

The term $|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \hat{\boldsymbol{\mu}}_k|_2^2$ comes from the strong convexity of the function H_{pen}. By simple algebra or using (3.58) with $\boldsymbol{g} = \hat{\boldsymbol{\mu}}_k$, we have

$$\mathrm{pen}(\hat{\boldsymbol{\theta}}_{\mathrm{pen}}) \quad + \quad \underbrace{|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \hat{\boldsymbol{\mu}}_k|_2^2}_{\substack{\text{Term given by the} \\ \text{strong convexity of } H_{\mathrm{pen}}}} \quad = \quad \sum_{j=1}^M \hat{\theta}_{\mathrm{pen},\,j} \quad \underbrace{|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2}_{\substack{\text{Term that controls} \\ \text{the deviations of } \Delta_{jk}}}. \tag{3.29}$$

Formula (3.29) highlights a feature of the penalty (3.8): the penalty transforms the quadratic term given by strong convexity into the linear term given by the right hand side of (3.29).

The strong convexity of $C_p(\cdot)$ and $H_{\mathrm{pen}}(\cdot)$ is understood with respect to the pseudometric

$$|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}'}|_2, \qquad \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^M,$$

so it is not the strong convexity in the Euclidean norm. We say that a function $V(\cdot)$ is strongly convex with coefficient $\gamma > 0$ over the simplex if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Lambda^M$,

$$V(\boldsymbol{\theta}) \geq V(\boldsymbol{\theta}') + \nabla V(\boldsymbol{\theta}')^T(\boldsymbol{\theta} - \boldsymbol{\theta}') + \gamma|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}'}|_2^2.$$

The strong convexity of H_{pen} could be used because H_{pen} is minimized over the simplex and not just over the vertices. Indeed, minimizing a strongly convex function over a discrete set, as in the definition of $\hat{J}$, only grants the inequalities

$$C_p(\boldsymbol{e}_{\hat{J}}) \leq C_p(\boldsymbol{e}_k), \qquad \text{for all } k = 1, ..., M.$$

Because the simplex is a convex set, minimizing the strongly convex function H_{pen} over the simplex grants the inequalities

$$H_{\mathrm{pen}}(\hat{\boldsymbol{\theta}}_{\mathrm{pen}}) \leq H_{\mathrm{pen}}(\boldsymbol{\theta}) - \frac{1}{2}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}}|_2^2, \qquad \text{for all } \boldsymbol{\theta} \in \Lambda^M.$$

One could also consider the estimator $\hat{\boldsymbol{\theta}}_C \in \mathrm{argmin}_{\boldsymbol{\theta}\in\Lambda^M} C_p(\boldsymbol{\theta})$. Because of the strong convexity of $C_p(\cdot)$, this estimator enjoys the inequalities

$$C_p(\hat{\boldsymbol{\theta}}_C) \leq C_p(\boldsymbol{\theta}) - |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}}|_2^2, \qquad \text{for all } \boldsymbol{\theta} \in \Lambda^M.$$

The above displays highlight the fact that $C_p(\cdot)$ and $H_{\mathrm{pen}}(\cdot)$ have different strong convexity coefficients. This is because $H_{\mathrm{pen}}(\cdot) = C_p(\cdot) + (1/2)\mathrm{pen}(\cdot)$ and $(1/2)\mathrm{pen}(\cdot)$ is strongly concave with coefficient $1/2$, thus the strong convexity coefficient of $H_{\mathrm{pen}}(\cdot)$ is less than that of $C_p(\cdot)$. We refer to Lemma 3.14 for a rigorous proof of the strong convexity of H_{pen} and C_p.

The estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_C}$ is another candidate for the problem of aggregation of affine estimators. It is close to the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}}$, except that the penalty (3.8) has been removed from the function to minimize. It was proved in [32, Section 2.2] that when

$A_j = 0$ for all $j = 1, ..., M$, this estimator performs poorly: for large enough M and n, there exist $\boldsymbol{f}$ and $\boldsymbol{b}_1, ..., \boldsymbol{b}_M \in \mathbb{R}^n$ such that with probability greater than $1/4$,

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_C} - \boldsymbol{f}|_2^2 \geq \min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + \frac{\sigma^2\sqrt{n}}{48},$$

where $\hat{\boldsymbol{\mu}}_j = \boldsymbol{b}_j$ for all $j = 1, ..., M$.

3.5 Prior weights

We consider now the problem of aggregation of M affine estimators with a prior probability distribution $\boldsymbol{\pi} = (\pi_1, ..., \pi_M)^T$ on the finite set of indices $\{1, ..., M\}$.

Theorem 3.6. *Let $M \geq 2$. For $j = 1, ..., M$, consider the estimator $\hat{\boldsymbol{\mu}}_j = A_j\mathbf{y} + \boldsymbol{b}_j$ and assume that $|||A_j|||_2 \leq 1$. Let $\boldsymbol{\pi} = (\pi_1, ..., \pi_M)^T \in \Lambda^M$. Assume that the noise $\boldsymbol{\xi}$ has distribution $\mathcal{N}(0, \sigma^2 I_{n\times n})$. Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}} \in \operatorname{argmin}_{\boldsymbol{\theta}\in\Lambda^M} V_{\mathrm{pen}}(\boldsymbol{\theta})$ where*

$$V_{\mathrm{pen}}(\boldsymbol{\theta}) := H_{\mathrm{pen}}(\boldsymbol{\theta}) + 30\sigma^2\mathcal{K}\theta. \tag{3.30}$$

Then for all $x > 0$, with probability greater than $1 - \exp(-x)$,

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} \left(|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 60\sigma^2 \log\frac{1}{\pi_j} \right) + 30\sigma^2 x. \tag{3.31}$$

Furthermore,

$$\mathbb{E}|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}} - \boldsymbol{f}|_2^2 \leq \mathbb{E} \min_{j=1,...,M} \left(|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 60\sigma^2 \log\frac{1}{\pi_j} \right). \tag{3.32}$$

The prior probability distribution $\boldsymbol{\pi} = (\pi_j)_{j=1,...,M}$ is deterministic and does not depend on the data $\mathbf{y} = (Y_1, ..., Y_n)^T$. The only difference between the function (3.7) and the function minimized in (3.30) is the term

$$\sigma^2\mathcal{K}\theta. \tag{3.33}$$

This term allows us to weight the candidates $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ with the prior probability distribution $(\pi_j)_{j=1,...,M}$ based on some prior knowledge about the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$. For example, if the estimators are orthoprojectors, one can set prior weights that decrease with the rank the orthoprojectors [92, 93]. The same term is used in [69] whereas [33] uses the Kullback-Leibler divergence of $\boldsymbol{\theta}$ from $\boldsymbol{\pi}$. It is shown in [32] that for aggregation of deterministic vectors, one may use a quantity of the form $\sum_{j=1}^M \theta_j \log(\rho(\theta_j)/\pi_j)$ where $\rho(\cdot)$ satisfies $\rho(t) \geq t$ and $t \to t\log(\rho(t))$ is convex. This suggests that we could use the Kullback-Leibler divergence of $\boldsymbol{\theta}$ from $\boldsymbol{\pi}$ instead of (3.33), but in their current form our proofs only hold with the "linear entropy" (3.33).

3.6 Robustness of the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}}$

We prove in this section that the procedure (3.9) is robust to non Gaussian noise distributions and to variance misspecification.

3.6.1 Robustness to non-Gaussian noise

The following result shows that the penalized procedure (3.9) is robust to non-Gaussian noise distributions.

Theorem 3.7. *Let $M \geq 2$. Let $\bar{\sigma} > 0$. For $j = 1, ..., M$, consider the estimator $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j$ and assume that $|||A_j|||_2 \leq 1$. Assume that the noise components $\xi_1, ..., \xi_n$ are i.i.d., centered with variance σ^2 and satisfy for all $\boldsymbol{b} \in \mathbb{R}^n$, all matrices B and all $x > 0$*

$$\mathbb{P}\left(\boldsymbol{\xi}^T \boldsymbol{b} > \bar{\sigma}\sqrt{2x}\right) \leq \exp(-x), \tag{3.34}$$

$$\mathbb{P}\left(\boldsymbol{\xi}^T B \boldsymbol{\xi} - \sigma^2 \mathrm{Tr} B > 2\sigma\bar{\sigma}\,|B|_{\mathrm{F}}\sqrt{x} + 2\bar{\sigma}^2 |||B|||_2 x\right) \leq \exp(-x). \tag{3.35}$$

Let $\hat{\boldsymbol{\theta}}_{\mathrm{pen}}$ be the estimator defined in (3.9). Then for all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}}$ satisfies with probability greater than $1 - 2\exp(-x)$,

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 46\bar{\sigma}^2(2\log M + x). \tag{3.36}$$

Let $K > 0$. If the random variables $\xi_1, ..., \xi_n$ are i.i.d., centered with variance σ^2 and K-subgaussian in the sense that $\log \mathbb{E}[e^{t\xi_i}] \leq K^2 t^2/2$ for all $t \in \mathbf{R}$ and all $i = 1, ..., n$, then (3.34) is satisfied with $\bar{\sigma} = cK$ for some absolute constant $c > 0$ [102, Section 5.2.3]. As $\sigma \leq K$, (3.34) is also satisfied with $\bar{\sigma} = cK^2/\sigma$. By the Hanson-Wright inequality [52, 105, 94], (3.35) also holds with $\bar{\sigma} = cK^2/\sigma$ for another absolute constant $c > 0$. Thus, for i.i.d. K-subgaussian random variables with variance σ^2, (3.36) yields

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + C(K^4/\sigma^2)(2\log M + x), \tag{3.37}$$

for some absolute constant $C > 0$. For most common examples of subgaussian random variables, the standard deviation σ is of the same order as the subgaussian norm K, so the bound (3.37) is satisfying. This bound may not be tight if the standard deviation is pathologically small compared to the subgaussian norm.

3.6.2 Robustness to variance misspecification

In order to construct the estimator (3.9) by minimizing (3.7), the knowledge of the variance of the noise is needed. However, the following proposition shows that the procedure (3.9) is robust to variance misspecification, i.e., the result still holds if the variance is replaced by an estimator $\hat{\sigma}^2$ as soon as $\hat{\sigma}^2$ is consistent in a weak sense defined below.

Theorem 3.8 (Aggregation under variance misspecification)**.** *Let $M \geq 2$. For $j = 1, ..., M$, consider the estimator $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j$. Assume that the noise random variables $\xi_1, ...\xi_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Let $\hat{\sigma}^2$ be an estimator and assume that*

$$\forall j = 1, ..., M,\ A_j = A_j^T = A_j^2, \qquad \delta := \mathbb{P}\left(|\sigma^2 - \hat{\sigma}^2| > \sigma^2/8\right) < 1. \tag{3.38}$$

Let $\hat{\boldsymbol{\theta}}_{\hat{\sigma}} = \mathrm{argmin}_{\boldsymbol{\theta} \in \Lambda^M} W_{\mathrm{pen}}(\boldsymbol{\theta})$ where

$$W_{\mathrm{pen}}(\boldsymbol{\theta}) := |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 - 2\mathbf{y}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + 2\hat{\sigma}^2 \mathrm{Tr}(A_{\boldsymbol{\theta}}) + \tfrac{1}{2}\mathrm{pen}(\boldsymbol{\theta}). \tag{3.39}$$

Then for all $x > 0$, with probability greater than $1 - \delta - \exp(-x)$,

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\hat{\sigma}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 48\sigma^2(x + 2\log M).$$

The proof of Theorem 3.8 is given in Section 3.8.3. In (3.38), the matrices $A_1, ..., A_M$ are assumed to be orthoprojectors, so Theorem 3.8 is a result for aggregation of Least Squares estimators. As soon as an estimator $\hat{\sigma}^2$ satisfies with high probability $|\hat{\sigma}^2 - \sigma^2| \leq \sigma^2/8$, optimal aggregation of Least Squares estimators is possible. This condition is weaker than consistency, as any estimator $\hat{\sigma}^2$ that converges to σ^2 in probability satisfies this condition for n large enough.

The proof of Theorem 3.8 exploits the form of the penalty (3.8) and the strong convexity of the function (3.39). Similarly to Proposition 3.5, we will prove that almost surely,

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\hat{\sigma}}} - \boldsymbol{f}|_2^2 \leq \min_{q=1,\dots,M} |\hat{\boldsymbol{\mu}}_q - \boldsymbol{f}|_2^2 + \max_{j,k=1,\dots,M} \left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2 + 2(\sigma^2 - \hat{\sigma}^2)\mathrm{Tr}(A_j - A_k) \right), \quad (3.40)$$

where Δ_{jk} is the quantity (3.19). The only difference from (3.24) is in the extra term $2(\sigma^2 - \hat{\sigma}^2)\mathrm{Tr}(A_j - A_k)$ that appears because we used $\hat{\sigma}^2$ instead of σ^2 in the definition of $W_{\mathrm{pen}}(\cdot)$. On the event $|\hat{\sigma}^2 - \sigma^2| \leq \sigma^2/8$, it is easy to check that (cf. Lemma 3.13)

$$2(\sigma^2 - \hat{\sigma}^2)\mathrm{Tr}(A_j - A_k) \leq \frac{\sigma^2}{4} \|A_j - A_k\|_{\mathrm{F}}^2 .$$

As explained in the discussion that follows Proposition 3.5, the quantity $\frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2$ is given by the interplay between the penalty (3.8) and the strong convexity of the function that is minimized. By (3.25), the expectation of this quantity is greater than $(\sigma^2/2)\|A_j - A_k\|_{\mathrm{F}}^2$. Thus, the penalty (3.8) and the strong convexity of W_{pen} provide exactly what is needed to compensate the difference between $\hat{\sigma}^2$ and σ^2. Hence, the proof of Theorem 3.8 reveals that the robustness to variance misspecification is in fact due to the interplay between the penalty (3.8) and the strong convexity of W_{pen}.

The papers [47, 49, 8] aim at performing aggregation of Least Squares estimators when σ^2 is unknown, but unlike Theorem 3.8 the oracle inequalities that they established have a leading constant greater than 1. To our knowledge, Theorem 3.8 is the first aggregation result, with leading constant 1, that is robust to variance misspecification.

In the following, we describe several situations where the suitable estimator $\hat{\sigma}^2$ is available.

Example 3.1 (An estimator $\hat{\sigma}^2$ that does not depend on $\mathbf{y}$)**.** In [35, Section 3.1], two contexts are given where an unbiased estimator of the covariance matrix, independent from $\mathbf{y}$, is available. For example, the noise level can be estimated independently if the signal is captured multiple times by a single device, or if several identical devices capture the same signal.

Example 3.2 (Difference based estimators)**.** In nonparametric regression where the non-random design points are equispaced in $[0, 1]$, a well known estimator of the noise level is the difference based estimator $1/(2n-2)\sum_{i=1}^{n-1}(y_{i+1} - y_i)^2$. This technique can be refined with more complex difference sequences [51, 39], and extends to design points in a multidimensional space [81]. For images, where the underlying space is 2-dimensional, there exist efficient methods which require no multiplication [56].

Example 3.3 (Consistent estimation of σ^2 in high-dimensional linear regression)**.** In a high-dimensional setting, it is possible to estimate σ^2 under classical assumptions

in high-dimensional regression. First, the scaled Lasso [95] allows a joint estimation of the regression coefficients and of the noise level σ^2. The estimator $\hat{\sigma}^2$ of the scaled Lasso converges in probability to the true noise level σ^2 [95, Theorem 1], and $\hat{\sigma}^2/\sigma^2$ is asymptotically normal [95, Equation (19)]. Second, [15] proposes to estimate σ^2 with with a recursive procedure that uses Lasso residuals, and non-asymptotic guarantees are proved [15, Supplementary material]. Third, [16] provides non-asymptotic bounds on the estimation of σ^2 by the residuals of the Square-Root Lasso [16, Theorem 2] and these bounds imply consistency. In Theorem 3.8, we require that $|\hat{\sigma}^2/\sigma^2 - 1| \leq 1/8$ with high-probability and this requirement is far weaker than the guarantees obtained in [15, 95].

3.7 Examples

3.7.1 Adaptation to the smoothness

For all $n \geq 1$, given continuous parameters $\beta \geq 1$ and $L > 0$, we consider subsets $\Theta(\beta, L) \subset \mathbb{R}^n$. We assume that for each $\beta \geq 1$, there exists a squared matrix A_β of size n with $|||A_\beta|||_2 \leq 1$ such that for all $L > 0$, as $n \to +\infty$,

$$\inf_{\hat{f}} \sup_{\boldsymbol{f} \in \Theta(\beta,L)} \frac{1}{n}\mathbb{E}|\boldsymbol{f} - \hat{\boldsymbol{f}}|_2^2 \sim \sup_{\boldsymbol{f} \in \Theta(\beta,L)} \frac{1}{n}\mathbb{E}|\boldsymbol{f} - A_\beta \mathbf{y}|_2^2 \sim C^* n^{\frac{-2\beta}{2\beta+1}}, \tag{3.41}$$

where $a_n \sim b_n$ if and only if $a_n/b_n \to 1$ as $n \to +\infty$, the infimum is taken over all estimators and the constant $C^* > 0$ may depend on β, L and σ. The above assumption holds for Sobolev ellipsoids in nonparametric regression, and in this case one can choose the Pinsker filters for the matrices A_β (cf. [99, Theorem 3.2]). For Sobolev ellipsoids, there exist different estimators that adapt to the unknown smoothness [43, 99, 35].

Consider the following aggregation procedure. Assume that $n \geq 3$ and let $M = \lceil 120 \log(n)(\log\log n)^2 \rceil$. For all $j = 1, ..., M$, let $\beta_j = (1 + 1/(\log(n)\log\log n))^{j-1}$. We aggregate the linear estimators $(\hat{\boldsymbol{\mu}}_j = A_{\beta_j}\mathbf{y})_{j=1,...,M}$ using the procedure (3.9) of Theorem 3.1, and denote by $\tilde{\boldsymbol{\mu}}$ the resulting estimator. The following adaptation result is a direct consequence of Theorem 3.1.

Proposition 3.9. *For all $n \geq 3$, $\beta \geq 1$ and $L > 0$, let $\Theta(\beta, L) \subset \mathbb{R}^n$ such that as $n \to +\infty$, (3.41) is satisfied for some matrices A_β with $|||A_\beta|||_2 \leq 1$. Assume that the sets $\Theta(\beta, L)$ are ordered, i.e., $\Theta(\beta, L) \subset \Theta(\beta', L)$ for any $\beta > \beta'$ and any $L > 0$. For all $\beta \geq 1$ and $L > 0$, the estimator $\tilde{\boldsymbol{\mu}}$ defined above satisfies as $n \to +\infty$*

$$\lim_{n\to+\infty} \sup_{\boldsymbol{f} \in \Theta(\beta,L)} \frac{1}{n}\mathbb{E}|\boldsymbol{f} - \tilde{\boldsymbol{\mu}}|_2^2 \; n^{\frac{2\beta}{2\beta+1}} = C^*.$$

The above procedure adapts to the unknown smoothness in exact asymptotic sense by aggregating only $\log(n)(\log\log n)^2$ estimators so its computational complexity is small. Another feature is that the minimax rate and the minimax constant C^* are not altered by the aggregation step.

3.7.2 The best convex combination as a benchmark

We consider convex combinations of the estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ to construct the estimator (3.9). The goal of this section is to study the performance of the estimator (3.9) if the benchmark is $\min_{\boldsymbol{\theta} \in \Lambda^M} |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2$ instead of $\min_{k=1,...,M} |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2$.

The penalty (3.8) vanishes at the extreme points: $\mathrm{pen}(\boldsymbol{e}_j) = 0$ for all $j = 1, ..., M$, and it pushes $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}}$ towards the points $\{\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M\}$. This can be seen in Figure 3.1. Consider a noise-free problem where $\sigma = 0$. Let $\boldsymbol{f} \in \mathbb{R}^n$. Consider estimators $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ such that $|\hat{\boldsymbol{\mu}}_j|_2^2 = \rho > 0$ for all $j = 1, ..., M$ (here, the estimators are deterministic because $\sigma = 0$). Then $\mathrm{pen}(\boldsymbol{\theta}) = \rho - |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2$ and $H_{\mathrm{pen}}(\boldsymbol{\theta}) = (1/2)|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - 2\boldsymbol{f}|_2^2 + c$ where c is constant that depends on ρ and $\boldsymbol{f}$ but not on $\boldsymbol{\theta}$. If both $\boldsymbol{f}$ and $2\boldsymbol{f}$ lie in the convex hull of $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$, $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}}$ defined in (3.9) will be equal to $2\boldsymbol{f}$ instead of $\boldsymbol{f}$ and is likely to be a bad procedure with respect to the benchmark $\min_{\boldsymbol{\theta} \in \Lambda^M} |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2$. This fact is not surprising since the penalty penalizes heavily some regions of the convex hull of the estimators. Furthermore this procedure is tailored for the benchmark $\min_{k=1,...,M} |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2$ and its goal is not to mimic the best convex combination of the estimators.

It is possible to modify the procedure (3.9) to construct an estimator that performs well with respect to the best convex combination of M linear estimators. Let

$$m \coloneqq \left\lfloor \sqrt{\frac{n}{\log(1 + M/\sqrt{n})}} \right\rfloor . \tag{3.42}$$

If $m \geq 1$, define the set $\Lambda_m^M \subset \Lambda^M$ as

$$\Lambda_m^M \coloneqq \left\{ \frac{1}{m} \sum_{q=1}^{m} \boldsymbol{u}_q, \quad \boldsymbol{u}_1, ..., \boldsymbol{u}_m \in \{\boldsymbol{e}_1, ..., \boldsymbol{e}_M\} \right\}. \tag{3.43}$$

Denote by $|\Lambda_m^M|$ the cardinality of Λ_m^M. We aggregate the affine estimators $(\hat{\boldsymbol{\mu}}_{\boldsymbol{u}})_{\boldsymbol{u} \in \Lambda_m^M}$ using the procedure (3.9) and denote by $\hat{\boldsymbol{\mu}}_{\Lambda_m^M}$ the resulting estimator.

Proposition 3.10. *Let $M, n \geq 1$. For $j = 1, ..., M$, consider the estimator $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j$ for any $n \times n$ matrix A_j and vector $\boldsymbol{b}_j \in \mathbb{R}^n$. Assume that $\boldsymbol{\xi} \sim N(0, \sigma^2 I_{n \times n})$ and that for some constant $R > 0$,*

$$\frac{1}{n}|\boldsymbol{f}|_2^2 \leq R^2, \qquad \frac{1}{n}|\boldsymbol{b}_j|_2^2 \leq R^2, \qquad |||A_j|||_2 \leq 1, \qquad \forall j = 1, ..., M.$$

For all $x > 0$, the estimator $\hat{\boldsymbol{\theta}}_C \in \mathrm{argmin}_{\boldsymbol{\theta} \in \Lambda^M} C_p(\boldsymbol{\theta})$ satisfies with probability greater than $1 - 2\exp(-x)$,

$$\frac{1}{n}|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_C} - \boldsymbol{f}|_2^2 \leq \min_{\boldsymbol{\theta} \in \Lambda^M} \frac{1}{n}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 + 8(\sigma^2 + \sigma R\sqrt{2})\sqrt{\frac{x + 2\log M}{n}} + \frac{8\sigma^2(x + 2\log M)}{n}. \tag{3.44}$$

If $M \leq \sqrt{n}(\exp(n) - 1)$ then for all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}_{\Lambda_m^M}$ defined above satisfies with probability greater than $1 - 3\exp(-x)$,

$$\frac{1}{n}|\hat{\boldsymbol{\mu}}_{\Lambda_m^M} - \boldsymbol{f}|_2^2 \leq \min_{\boldsymbol{\theta} \in \Lambda^M} \frac{1}{n}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 + C\max(R^2, \sigma^2)\sqrt{\frac{\log(1 + M/\sqrt{n})}{n}} + \frac{C\sigma^2 x}{n}. \tag{3.45}$$

To our knowledge, this is the first result that provides a sharp oracle inequality for the problem of aggregation of affine estimators with respect to the convex oracle. However, there is a large literature on convex aggregation when the estimators to aggregate are deterministic, which corresponds to the particular case $A_j = 0$ for all $j = 1, ..., M$. When the error is measured with the scaled squared norm $\frac{1}{n}|\cdot|_2^2$, the minimax rate of convex aggregation is known to be of order M/n if $M \leq \sqrt{n}$

and $\sqrt{\log(1+M/\sqrt{n})/n}$ if $M > \sqrt{n}$. For our setting, this is proved in [92]. This elbow effect was first established for regression with random design [98] and then extended to other settings in [89, 90]. All these results assume that the estimators to aggregate are deterministic or independent of the data used for aggregation. The lower bound [92, Theorem 5.3 with $S = M$, $\delta = \sigma$ and $R = \log(1 + eM)$] yields that there exist absolute constants $c, C > 0$ such that if $\log(1 + eM)^2 \le Cn$, there exist deterministic vectors $\hat{\boldsymbol{\mu}}_1 = \boldsymbol{b}_1, ..., \hat{\boldsymbol{\mu}}_M = \boldsymbol{b}_M$ such that for all estimators $\hat{\boldsymbol{\mu}}$,

$$\sup_{\boldsymbol{f}\in\mathbb{R}^n} \mathbb{P}_{\boldsymbol{f}} \left(\tfrac{1}{n}|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 - \min_{\boldsymbol{\theta}\in\Lambda^M} \tfrac{1}{n}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 \ge c\sigma^2 \left(\frac{M}{n} \wedge \sqrt{\frac{\log(1+M/\sqrt{n})}{n}} \right) \right) \ge c.$$

Thus, if $M \ge \sqrt{n}$, (3.45) is optimal in a minimax sense up to absolute constants, and (3.44) is optimal up to logarithmic factors. However, we do not know whether the minimax rate is M/n when $M < \sqrt{n}$, as in the case of aggregation of deterministic vectors.

The problem of linear aggregation of affine estimators remains open. It is only known that for linear aggregation of deterministic vectors, the Least Squares estimator on a linear space of dimension M achieves the rate $\sigma^2 M/n$, which is optimal in a minimax sense [98, 89, 92, 90].

3.7.3 k-regressors

Let $\mathbb{X}$ be a design matrix consisting of p columns and n rows. Let k be an integer such that $1 \le k \le n$. Consider the family of distinct estimators $\{\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M\}$ where for each $j = 1, ..., M$, $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y}$ and A_j is the orthoprojector on a linear span of k linearly independent columns of $\mathbb{X}$. In particular, $M \le \binom{p}{k}$. The estimator $\hat{\boldsymbol{\mu}}_j$ is the Least Squares estimator on the subspace V_j of dimension k which is the linear span of these k columns.

Now consider the estimator $\hat{\boldsymbol{\theta}}^{(k)} \in \mathbb{R}^M$ defined by

$$\hat{\boldsymbol{\theta}}^{(k)} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Lambda^M} \left(|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{y}|_2^2 + \frac{1}{2}\mathrm{pen}(\boldsymbol{\theta}) \right),$$

where $\mathrm{pen}(\cdot)$ is the penalty (3.8). It is exactly the procedure (3.9) from Theorem 3.1 since the projection matrices $A_1, ..., A_M$ have the same trace equal to k. This procedure is fully adaptive with respect to the unknown variance of the noise. The following result is a direct consequence of Theorem 3.1. The estimator $\hat{\boldsymbol{\theta}}^{(k)}$ satisfies for all $x > 0$, with probability greater than $1 - 3\exp(-x)$,

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}^{(k)}} - \boldsymbol{f}|_2^2 \le \min_{\boldsymbol{\beta}\in\mathbf{R}^p,\, |\boldsymbol{\beta}|_0\le k} |\mathbb{X}\boldsymbol{\beta} - \boldsymbol{f}|_2^2 + c\sigma^2 \left(k \log\left(\frac{ep}{k}\right) + x \right),$$

for some absolute constant $c > 0$.

3.7.4 Sparsity pattern aggregation

Given a design matrix $\mathbb{X}$ with p columns, an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{f}$ is said to achieve a sparsity oracle inequality if it satisfies

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 \le \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(C|\mathbb{X}\boldsymbol{\theta} - \boldsymbol{f}|_2^2 + \Delta(|\theta|_0) \right), \tag{3.46}$$

with high probability or in expectation. In (3.46), $C \geq 1$, $|\theta|_0$ is the number of non-zero coefficients of θ and Δ is an increasing function, which may also depend on problem parameters such as the variance of the noise or the design matrix $\mathbb{X}$. See (3.47) below for a typical example of such function $\Delta(\cdot)$. Results of the form (3.46) are of major interest in high-dimensional statistics where the number of covariates p exceeds the number of observations [92]. First approach to get such results can be found in [41], and in an expanded form in [18, 19], under Gaussian noise with a leading constant $C > 1$. The drawback of having $C > 1$ cannot be repaired for these penalized model selection procedures (cf. (3.23) and [46, Section 6.4.2]). More recently, aggregation methods based on exponential weights [92, 93, 96] and then Q-aggregation [33] were shown to achieve sharp oracle inequalities similar to (3.46). These sharp oracle inequalities were proved for Gaussian noise with known variance.

Aggregation procedures with prior weights [36, 32, 69, 12] as in Theorem 3.6 can be used to prove sparsity oracle inequalities if sparsity-inducing prior weights are used. For instance, sparsity pattern aggregation [92, 93, 33, 96] leads to the following oracle inequality. Given a design matrix $\mathbb{X}$ with p columns, there exists an estimator $\hat{\boldsymbol{\mu}}$ that satisfies with probability greater than $1 - 2\exp(-x)$,

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 \leq \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left(|\mathbb{X}\boldsymbol{\theta} - \boldsymbol{f}|_2^2 + c\,\sigma^2 |\boldsymbol{\theta}|_0 \log\left(\frac{ep}{1 \vee |\boldsymbol{\theta}|_0} \right) \right) + c'\sigma^2 x, \tag{3.47}$$

where $c, c' > 0$ are absolute constants and $|\boldsymbol{\theta}|_0$ denotes the number of non-zero coefficients of $\boldsymbol{\theta}$. When the noise is Gaussian, the result (3.47) is proved in [33] and a similar result in expectation was shown in [92, 96].

We now derive a similar result for subgaussian noise. We propose below a new sparsity pattern aggregation method that only requires an estimator $\hat{K}^2$ that upper bounds the subgaussian norm of the noise with high probability. We will make the following assumption on the noise. For some constant $K > 0$, we assume that the random vector $\boldsymbol{\xi}$ satisfies:

$$\forall \alpha \in \mathbb{R}^n, \qquad \mathbb{E}\exp(\alpha^T \boldsymbol{\xi}) \leq \exp\left(\frac{|\alpha|_2^2 K^2}{2} \right). \tag{3.48}$$

As opposed to the previous section, the components of $\boldsymbol{\xi}$ are not assumed to be independent.

For each subset $J \subset \{1, ..., p\}$, let $\hat{\boldsymbol{\mu}}_J^{LS}$ be the Least Squares estimator on the linear span of the columns of $\mathbb{X}$ whose indices are in J. The estimator $\hat{\boldsymbol{\mu}}_J^{LS}$ is of the form $\hat{\boldsymbol{\mu}}_J^{LS} = A_J \mathbf{y}$ for some projection matrix A_J. Consider the weights $\pi_J \propto e^{-|J|} \binom{p}{|J|}^{-1}$ and choose the normalisation constant such that $\sum_{J \subseteq \{1,...,p\}} \pi_J = 1$. Given $\boldsymbol{\lambda} = (\lambda_J)_{J \subseteq \{1,...,p\}}$, let $\hat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}} = \sum_{J \subseteq \{1,...,p\}} \lambda_J A_J \mathbf{y}$. Let $\hat{\boldsymbol{\mu}}_{\text{SPA}} = \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\lambda}}}$ where $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_J)_{J \subseteq \{1,...,p\}}$ is a minimizer of

$$|\mathbf{y} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}}|_2^2 + \sum_{J \subseteq \{1,...,p\}} \lambda_J \left(\frac{1}{2} |A_J \mathbf{y} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}}|_2^2 + 32\hat{K}^2 \log \frac{1}{\pi_J} \right)$$

over the set

$$\Lambda = \left\{ \boldsymbol{\lambda} = (\lambda_J)_{J \subseteq \{1,...,p\}}, \quad \sum_{J \subseteq \{1,...,p\}} \lambda_J = 1, \quad \lambda_J \geq 0, \forall J \subseteq \{1, ..., p\} \right\}.$$

As sparsity pattern aggregation is not central in the present paper, we keep this presentation short and refer the reader to [92, 93, 33, 96] for more details on sparsity pattern aggregation and the construction of Least Squares estimators.

Then the following sparsity oracle inequality holds, where $|\boldsymbol{\theta}|_0$ is the number of non-zero coefficients of $\boldsymbol{\theta}$.

Theorem 3.11. *Let $\mathbb{X}$ be a deterministic design matrix with p columns Let $K > 0$ be the smallest positive real number such that the noise random $\boldsymbol{\xi}$ satisfies (3.48). Let $\hat{K}$ be a given estimator and let $\delta := \mathbb{P}(\hat{K}^2 < K^2)$. Then, the estimator $\hat{\boldsymbol{\mu}}_{\text{SPA}}$ defined above satisfies with probability greater than $1 - \delta - 3\exp(-x)$,*

$$|\hat{\boldsymbol{\mu}}_{\text{SPA}} - \boldsymbol{f}|_2^2 \le \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \Bigg[|\mathbb{X}\boldsymbol{\theta} - \boldsymbol{f}|_2^2 + 31K^2 x + (64\hat{K}^2 + 4K^2)\left(\frac{1}{2} + 2|\boldsymbol{\theta}|_0 \log\left(\frac{ep}{1 \vee |\boldsymbol{\theta}|_0}\right)\right)\Bigg]. \quad (3.49)$$

Theorem 3.11 is proved in the supplementary material. It improves upon the previous results on sparsity pattern aggregation [33, 92, 93, 96] in several aspects.

First, the noise $\boldsymbol{\xi}$ is only assumed to be subgaussian and its components need not be independent, whereas previous results only hold under Gaussianity and independence of the noise components [33, 92, 93, 96]. Theorem 3.11 shows that the optimal bounds are of the same form in this more general setting.

Second, to construct the aggregates in [33, 92, 93, 96] one needs the exact knowledge of the covariance matrix of the noise. In Theorem 3.11, only an upper bound of the subgaussian norm of the noise is needed to construct the estimator.

Third, we do not split the data in order to perform sparsity pattern aggregation, as opposed to the "sample cloning" approach [96, Lemma 2.1]. Sample cloning is possible only for Gaussian noise when the variance is known; it cannot be used here as $\boldsymbol{\xi}$ can be any subgaussian vector.

The estimator of Theorem 3.11 achieves the minimax rate for any intersection of ℓ_0 and ℓ_q balls, where $q \in (0, 2)$. This can be shown by applying the arguments of [33, 96] and bounding the right hand side of (3.49). Indeed, although [33, 96] consider only normal random variables, the argument does not depend on the noise distribution.

The result above holds without any assumption on the design matrix $\mathbb{X}$, as opposed to the Lasso or the Dantzig estimators which need assumptions on the design matrix $\mathbb{X}$ to achieve sparsity oracle inequalities similar but weaker than (3.49).

The interest of the Lasso and the Dantzig estimators is that they can be computed efficiently for large p. The sparsity pattern aggregate based on exponential weights can also be computed efficiently usind MCMC methods [92]. The estimator $\hat{\theta}^{SPA}$ proposed here suffers the same drawback as [18] or the sparsity pattern aggregate performed with Q-aggregation [33]: it is not known whether these estimators can be computed in polynomial time, which makes them useful only for relatively small p.

3.8 Proofs

3.8.1 Preliminaries

The following notation will be useful. Define for all $j, k = 1, ..., M$

$$Q_{j,k} \coloneqq \left(-2I_{n\times n} - \tfrac{1}{2}(A_k - A_j)^T\right)(A_k - A_j), \tag{3.50}$$

$$\boldsymbol{v}_{j,k} \coloneqq \left(-2I_{n\times n} - (A_k - A_j)^T\right)((A_k - A_j)\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j)\,. \tag{3.51}$$

Let $B_{jk} = A_k - A_j$, so that $\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j = B_{jk}\boldsymbol{\xi} + (B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j)$. Then

$$|\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j|_2^2 = |B_{jk}\boldsymbol{\xi}|_2^2 + |B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2^2 + 2\boldsymbol{\xi}^T B_{jk}^T(B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j).$$

Thus, simple algebra yields that the quantity Δ_{jk} defined in (3.19) satisfies

$$\begin{aligned}\Delta_{jk} - \tfrac{1}{2}|\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j|_2^2 = \boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi}] + \boldsymbol{\xi}^T\boldsymbol{v}_{j,k} \\ - \frac{\sigma^2}{2}\left\|A_j - A_k\right\|_{\mathrm{F}}^2 - \frac{1}{2}|(A_k - A_j)\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2^2,\end{aligned} \tag{3.52}$$

where we used the equality $\sigma^2\left\|A_j - A_k\right\|_{\mathrm{F}}^2 = \mathbb{E}[|(A_j - A_k)\boldsymbol{\xi}|_2^2]$ and the above definitions of $Q_{j,k}$ and $\boldsymbol{v}_{j,k}$. Furthermore, using (3.5) and $\left|\!\left|\!\left|A_j - A_k\right|\!\right|\!\right|_2 \le 2$ we have

$$\begin{aligned}\left|\!\left|\!\left|Q_{j,k}\right|\!\right|\!\right|_2 \le 6, \qquad \left\|Q_{j,k}\right\|_{\mathrm{F}} \le 3\left\|A_k - A_j\right\|_{\mathrm{F}}, \\ |\boldsymbol{v}_{j,k}|_2 \le 4|(A_k - A_j)\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2\end{aligned} \tag{3.53}$$

for all $j, k = 1, ..., M$. This yields that

$$\begin{aligned}\Delta_{jk} - \tfrac{1}{2}|\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j|_2^2 \le \boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi}] + \boldsymbol{\xi}^T\boldsymbol{v}_{j,k} \\ - \frac{\sigma^2}{18}\left\|Q_{j,k}\right\|_{\mathrm{F}}^2 - \frac{1}{32}|\boldsymbol{v}_{j,k}|_2^2.\end{aligned} \tag{3.54}$$

Proposition 3.12. *Let $\boldsymbol{v} \in \mathbb{R}^n$ and let Q be any squared matrix of size n. Assume that $\xi_1, ..., \xi_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Then for all $u > 0$ such that $2u\sigma^2\left|\!\left|\!\left|Q\right|\!\right|\!\right|_2 < 1$ we have*

$$\mathbb{E}\left[e^{u\left(\boldsymbol{\xi}^T Q\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q\boldsymbol{\xi}] + \boldsymbol{\xi}^T\boldsymbol{v}\right)}\right] \le \exp\left(u^2\sigma^2\left(\frac{\sigma^2\left\|Q\right\|_{\mathrm{F}}^2 + \frac{|\boldsymbol{v}|_2^2}{2}}{1 - 2\sigma^2\left|\!\left|\!\left|Q\right|\!\right|\!\right|_2 u}\right)\right). \tag{3.55}$$

Furthermore, define

$$\begin{aligned}Z_{Q,\boldsymbol{v}} &\coloneqq \boldsymbol{\xi}^T Q\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q\boldsymbol{\xi}] + \boldsymbol{\xi}^T\boldsymbol{v} - \frac{\sigma^2}{18}\left\|Q\right\|_{\mathrm{F}}^2 - \frac{1}{32}|\boldsymbol{v}|_2^2, \\ Y_{Q,\boldsymbol{v}} &\coloneqq \boldsymbol{\xi}^T Q\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q\boldsymbol{\xi}] + \boldsymbol{\xi}^T\boldsymbol{v} - \frac{\sigma^2}{36}\left\|Q\right\|_{\mathrm{F}}^2 - \frac{1}{32}|\boldsymbol{v}|_2^2.\end{aligned}$$

If $\left|\!\left|\!\left|Q\right|\!\right|\!\right|_2 \le 6$, then for $u = 1/(30\sigma^2)$ we have $\mathbb{E}\left[e^{uZ_{Q,\boldsymbol{v}}}\right] \le 1$ and for $u' = 1/(48\sigma^2)$ we have $\mathbb{E}\left[e^{u'Y_{Q,\boldsymbol{v}}}\right] \le 1$.

The proof relies on an argument similar to that of [63, Lemma 1].

Proof. If Q is not symmetric, let $Q_s = (Q + Q^T)/2$. We have $\|Q_s\|_{\mathrm{F}} \le \|Q\|_{\mathrm{F}}$, $|||Q_s|||_2 \le |||Q|||_2$ and almost surely $\boldsymbol{\xi}^T Q \boldsymbol{\xi} = \boldsymbol{\xi}^T Q_s \boldsymbol{\xi}$ so that if (3.55) holds for Q_s then

$$\mathbb{E}\left[e^{u\left(\boldsymbol{\xi}^T Q \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q \boldsymbol{\xi}] + \boldsymbol{\xi}^T \boldsymbol{v}\right)}\right] \le e^{u^2\sigma^2\left(\frac{\sigma^2\|Q_s\|_{\mathrm{F}}^2 + \frac{|\boldsymbol{v}|_2^2}{2}}{1 - 2\sigma^2|||Q_s|||_2 u}\right)} \le e^{u^2\sigma^2\left(\frac{\sigma^2\|Q\|_{\mathrm{F}}^2 + \frac{|\boldsymbol{v}|_2^2}{2}}{1 - 2\sigma^2|||Q|||_2 u}\right)}.$$

Thus the result for the symmetric matrix Q_s implies the result for Q.

We now assume that Q is symmetric. There exists a matrix P with $P^T P = P P^T = I_{n\times n}$ such that $Q = P^T \mathrm{diag}(\lambda_1, ..., \lambda_n) P$, where $\lambda_1, ..., \lambda_n$ are the eigenvalues of Q. Let $\boldsymbol{w} = (1/\sigma) P \boldsymbol{v}$ and define the random variables $g_1, ..., g_n$ by $(g_1, ..., g_n)^T = (1/\sigma) P \boldsymbol{\xi}$. By the rotational invariance of the Gaussian distribution, $g_1, ..., g_n$ are i.i.d. $\mathcal{N}(0,1)$ random variables. Thus, the random variable $\boldsymbol{\xi}^T Q \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q \boldsymbol{\xi}] + \boldsymbol{\xi}^T \boldsymbol{v}$ has the same distribution as

$$\sigma^2 \sum_{i=1}^n W_i, \qquad \text{where} \qquad W_i := \lambda_i (g_i^2 - 1) + g_i w_i,$$

For all $i = 1, ..., n$ and for all $t > 0$ such that $\max_{i=1,...,n} 2t|\lambda_i| < 1$, integration using the probability density function of g_i yields

$$\mathbb{E}[e^{tW_i}] = \frac{1}{\sqrt{1 - 2\lambda_i t}} e^{\frac{t^2 w_i^2}{2(1-2\lambda_i t)} - t\lambda_i} \le e^{\frac{\lambda_i^2 t^2}{1 - 2|\lambda_i| t} + \frac{t^2 w_i^2}{2(1-2\lambda_i t)}},$$

where we used the inequalities

$$\log\left(\frac{1}{\sqrt{1-2v}}\right) \le v + \frac{v^2}{1-2v} = v + \frac{v^2}{1-2|v|} \qquad \text{for all } v \in [0, 1/2),$$
$$\log\left(\frac{1}{\sqrt{1-2v}}\right) \le v + v^2 \le v + \frac{v^2}{1-2|v|} \qquad \text{for all } v \in (-1/2, 0].$$

This can be shown by comparing the power series expansions. As $|\lambda_i| \le |||Q|||_2$ for all $i = 1, ..., n$, by independence of $W_1, ..., W_n$ we obtain

$$\mathbb{E}[e^{t\sum_{i=1}^n W_i}] \le \exp\left(t^2 \left(\frac{\|Q\|_{\mathrm{F}}^2 + \frac{|\boldsymbol{w}|_2^2}{2}}{1 - 2|||Q|||_2 t}\right)\right).$$

By definition of $\boldsymbol{w}$ we have $|\boldsymbol{v}|_2 = \sigma|\boldsymbol{w}|_2$, so setting $t = u\sigma^2$ completes the proof of (3.55).

The claims about $Z_{Q,\boldsymbol{v}}$ and $Y_{Q,\boldsymbol{v}}$ are direct consequences of (3.55). □

3.8.2 Proof of the main results

Proof of Theorem 3.1. By (3.5), it is enough to prove that

$$D_1 := \mathbb{E}\left[e^{u \max_{j,k=1,...,M}(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2)}\right] \le M^2 \tag{3.56}$$

for $u = 1/(30\sigma^2)$. Then, Jensen's inequality yields (3.13) and a Chernoff bound yields (3.12).

We now prove (3.56). By (3.53), for all $j, k = 1, ..., M$ we have $|||Q_{j,k}|||_2 \le 6$. Using (3.54) and Proposition 3.12, we have

$$D_1 \le \sum_{j=1}^M \sum_{k=1}^M \mathbb{E}\left[e^{u(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2)}\right] \le \sum_{j=1}^M \sum_{k=1}^M \mathbb{E}\left[e^{uZ_{Q_{j,k}, \boldsymbol{v}_{j,k}}}\right] \le M^2,$$

where for any matrix Q and any $\boldsymbol{v} \in \mathbb{R}^n$, the random variable $Z_{Q,\boldsymbol{v}}$ is defined in Proposition 3.12. □

Proof of Theorem 3.6. Let $\beta = 30\sigma^2$. Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}$ for notational simplicity. The only difference between H_{pen} and V_{pen} is the linear term (3.33). As in the proof of (3.24) in Section 3.4, by convexity of V_{pen} we have that for all $k = 1, ..., M$,

$$\begin{aligned}
& |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 \\
\leq\ & |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \nabla V_{\text{pen}}(\hat{\boldsymbol{\theta}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}), \\
=\ & 2\beta \log \frac{1}{\pi_k} + \sum_{j=1}^{M} \hat{\theta}_j \left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2 - \beta \log \frac{1}{\pi_j \pi_k} \right), \\
\leq\ & 2\beta \log \frac{1}{\pi_k} + \max_{j=1,...,M} \left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2 - \beta \log \frac{1}{\pi_j \pi_k} \right),
\end{aligned}$$

where Δ_{jk} is defined in (3.19). For all $u > 0$, let

$$D_2 := \mathbb{E}\left[\exp\left(u \max_{j,k=1,...,M} \left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2 - \beta \log \frac{1}{\pi_j \pi_k} \right) \right) \right].$$

We now bound from above this moment generating function using (3.54) and Proposition 3.12. If $u = 1/\beta = 1/(30\sigma^2)$ then

$$\begin{aligned}
D_2 &\leq \sum_{j=1}^{M} \sum_{k=1}^{M} \pi_j \pi_k \mathbb{E}\left[e^{u(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2)} \right] \\
&\leq \sum_{j=1}^{M} \sum_{k=1}^{M} \pi_j \pi_k \mathbb{E}\left[e^{u Z_{Q_{j,k}, \boldsymbol{v}_{j,k}}} \right] \leq \sum_{j=1}^{M} \sum_{k=1}^{M} \pi_j \pi_k = 1.
\end{aligned}$$

As in the proof of Theorem 3.1, Jensen's inequality yields (3.31) while a Chernoff bound completes the proof of (3.31). □

Proof of Theorem 3.7. For a fixed pair (j, k), we apply (3.34) to the vector $\boldsymbol{v}_{j,k}$ and (3.35) to the matrix $Q_{j,k}$. Using (3.53),

$$\begin{aligned}
\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k} \boldsymbol{\xi}] &\leq \bar{\sigma}^2 12x + 6\sigma\bar{\sigma} \left\| A_k - A_j \right\|_{\text{F}} \sqrt{x}, \\
&\leq 30\bar{\sigma}^2 x + \frac{\sigma^2}{2} \left\| A_k - A_j \right\|_{\text{F}}^2, \\
\boldsymbol{\xi}^T \boldsymbol{v}_{j,k} &\leq \bar{\sigma} 4 |(A_k - A_j)\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2 \sqrt{2x}, \\
&\leq 16\bar{\sigma}^2 x + \frac{1}{2}|(A_k - A_j)\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2^2.
\end{aligned}$$

Combining this bound with (3.40), (3.52) and the union bound completes the proof. □

3.8.3 Proof of Theorem 3.8

The following inequality will be useful.

Lemma 3.13 (Projection matrices)**.** *Let A, B be two squared matrices of size n with $A^T = A = A^2$ and $B^T = B = B^2$. Then*

$$|\text{Tr}(A - B)| \leq \|A - B\|_{\text{F}}^2. \tag{3.57}$$

Proof. Without loss of generality, assume that $\mathrm{Tr}A \geq \mathrm{Tr}B$. As $\|A-B\|_{\mathrm{F}}^2 = \|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2 - 2\mathrm{Tr}(AB)$ and $\|A\|_{\mathrm{F}}^2 = \mathrm{Tr}A$, (3.57) is equivalent to $2\mathrm{Tr}(AB) \leq 2\mathrm{Tr}(B)$. Notice that for projection matrices, $\mathrm{Tr}(AB) = \|AB\|_{\mathrm{F}}^2 \leq \|\|A\|\|_2^2 \|B\|_{\mathrm{F}}^2 \leq \|B\|_{\mathrm{F}}^2 = \mathrm{Tr}(B)$ and the proof is complete. □

Proof of Theorem 3.8. Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\hat{\sigma}}$ for notational simplicity. As in the proof of (3.24) in Section 3.4, by convexity of W_{pen} we have that for all $k = 1, ..., M$,

$$\begin{aligned}
& |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 \\
\leq\ & |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \nabla W_{\mathrm{pen}}(\hat{\boldsymbol{\theta}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}), \\
=\ & \sum_{j=1}^M \hat{\theta}_j \left(\Delta_{jk} + 2(\hat{\sigma}^2 - \sigma^2)\mathrm{Tr}(A_j - A_k) - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2 \right), \\
\leq\ & \max_{j=1,...,M} \left(\Delta_{jk} + 2(\hat{\sigma}^2 - \sigma^2)\mathrm{Tr}(A_j - A_k) - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2 \right) =: D_3,
\end{aligned}$$

where Δ_{jk} is defined in (3.19). The assumption on $\hat{\sigma}^2$ and (3.57) yield that on an event Ω_0 of probability greater than $1-\delta$,

$$2\left|(\hat{\sigma}^2 - \sigma^2)\mathrm{Tr}(A_j - A_k)\right| \leq \frac{\sigma^2}{4}\|A_j - A_k\|_{\mathrm{F}}^2 \qquad \text{for all } j,k = 1, ..., M.$$

Using (3.52) and (3.53), we obtain that on the event Ω_0,

$$\begin{aligned}
D_3 &\leq \max_{j,k=1,...,M} \left(\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^T Q_{j,k}\boldsymbol{\xi}] + \boldsymbol{\xi}^T \boldsymbol{v}_{j,k} - \frac{\sigma^2}{36}\|Q_{j,k}\|_{\mathrm{F}}^2 - \frac{1}{32}|\boldsymbol{v}_{j,k}|_2^2 \right), \\
&= \max_{j,k=1,...,M} Y_{Q_{j,k},\boldsymbol{v}_{j,k}},
\end{aligned}$$

where $Q_{j,k}$ and $\boldsymbol{v}_{j,k}$ are defined in (3.50) and (3.51) while $Y_{Q,\boldsymbol{v}}$ is defined in Proposition 3.12 for any matrix Q and any $\boldsymbol{v} \in \mathbb{R}^n$. Using Proposition 3.12, for $u = 1/(48\sigma^2)$ we have

$$\mathbb{E}\left[\exp\left(u \max_{j,k=1,...,M} Y_{Q_{j,k},\boldsymbol{v}_{j,k}}\right)\right] \leq \sum_{j=1}^M \sum_{k=1}^M \mathbb{E}\left[\exp\left(uY_{Q_{j,k},\boldsymbol{v}_{j,k}}\right)\right] \leq M^2.$$

By a Chernoff bound, this proves that on an event Ω_1 of probability greater than $1 - e^{-x}$, we have $\max_{j,k=1,...,M} Y_{Q_{j,k},\boldsymbol{v}_{j,k}}, \leq 48\sigma^2(x + 2\log M)$. On the event $\Omega_0 \cap \Omega_1$ we have $D_3 \leq 48\sigma^2(x + 2\log M)$ and the union bound yields that $\mathbb{P}(\Omega_0 \cap \Omega_1) \geq 1 - e^{-x} - \delta$. □

3.8.4 Strong convexity

The penalty (3.8) satisfies for any $\boldsymbol{g} \in \mathbb{R}^n$ and any $\boldsymbol{\theta} \in \Lambda^M$:

$$\sum_{k=1}^M \theta_k |\hat{\boldsymbol{\mu}}_k - \boldsymbol{g}|_2^2 = |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{g}|_2^2 + \mathrm{pen}(\boldsymbol{\theta}). \tag{3.58}$$

This can be shown by using simple properties of the Euclidean norm, or by noting that the equality above is a bias-variance decomposition. For $\boldsymbol{g} = 0$, (3.58) yields $\mathrm{pen}(\boldsymbol{\theta}) = -|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 + \sum_{k=1}^M \theta_k |\hat{\boldsymbol{\mu}}_k|_2^2$.

Lemma 3.14. *Let F be any one of the functions H_{pen}, V_{pen}, W_{pen} or U defined in (3.7), (3.30), (3.39) and the supplementary material, respectively. Then F is convex, differentiable and satisfies for all $\boldsymbol{\theta}, \boldsymbol{\theta}_0 \in \Lambda^M$,*

$$F(\boldsymbol{\theta}) = F(\boldsymbol{\theta}_0) + \nabla F(\boldsymbol{\theta}_0)^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_0}|_2^2. \tag{3.59}$$

Furthermore, if $\hat{\boldsymbol{\theta}}$ is a minimizer of F over the simplex then for all $\boldsymbol{\theta} \in \Lambda^M$,

$$F(\boldsymbol{\theta}) \geq F(\hat{\boldsymbol{\theta}}) + \frac{1}{2}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}}|_2^2. \tag{3.60}$$

Proof. Using (3.58) with $\boldsymbol{g} = 0$ we obtain that the function F is a polynomial of degree 2, of the form $F(\boldsymbol{\theta}) = \text{affline}(\boldsymbol{\theta}) + \frac{1}{2}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2$ where affine$(\cdot)$ is an affine function of $\boldsymbol{\theta}$. This shows that F is convex and differentiable. The result (3.59) follows by uniqueness of the Taylor expansion of F (or by an explicit calculation of $\nabla F(\boldsymbol{\theta}_0)$). Inequality (3.60) is a consequence of [21, 4.2.3, equation (4.21)]. □

3.8.5 Lower bound

Proof of Proposition 3.2. The lower bounds of [92, Theorem 5.4] are stated in expectation, but inspection of the proof of [92, Theorem 5.3 with $S = 1$, $\delta = \infty$ and $R = \log(1 + eM)$] reveals that the lower bound holds also in probability since it is an application of [99, Theorem 2.7]. This result yields that there exist absolute constants $p, c, C > 0$ and $\boldsymbol{f}_1, ..., \boldsymbol{f}_M \in \mathbb{R}^n$ such that for any estimator $\hat{\boldsymbol{\mu}}$,

$$\sup_{j=1,...,M} \mathbb{P}_{\boldsymbol{f}_j}(\Omega_j) \geq p, \qquad \Omega_j := \left\{|\hat{\boldsymbol{\mu}} - \boldsymbol{f}_j|_2^2 \geq c\sigma^2 \log(M)\right\},$$

provided that $\log(M) \leq cn$ and $n, M > C$. Set $\boldsymbol{b}_j = \boldsymbol{f}_j$ for all $j = 1, ..., M$. This lower bound implies that for any estimator $\hat{\boldsymbol{\mu}}$,

$$\sup_{\boldsymbol{f} \in \mathbb{R}^n} \mathbb{P}_{\boldsymbol{f}}\left(|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 - \min_{k=1,...,M} |\boldsymbol{b}_k - \boldsymbol{f}|_2^2 \geq c\sigma^2 \log(M)\right) \geq p.$$

For all $j = 1, ..., M$, let $A_j = (1/|\boldsymbol{f}_j|_2^2)\boldsymbol{f}_j\boldsymbol{f}_j^T$ so that A_j is the orthogonal projection on the linear span of $\boldsymbol{f}_j$. The orthoprojector A_j has rank one so under $\mathbb{P}_{\boldsymbol{f}_j}$, $|A_j\mathbf{y} - \boldsymbol{f}_j|_2^2/\sigma^2$ is a χ^2 random variable with one degree of freedom. Let Ω'_j be the event $\{|A_j\mathbf{y} - \boldsymbol{f}_j|_2^2 \leq c\sigma^2 \log(M)/2\}$ and let $\bar{\Omega}'_j$ be its complementary event. A two sided bound on the Gaussian tail implies that $\mathbb{P}_{\boldsymbol{f}_j}(\bar{\Omega}'_j) \leq 2/(M^{c/4})$, which is smaller than $p/2$ if M is larger than some absolute constant, so that we have $\mathbb{P}_{\boldsymbol{f}_j}(\bar{\Omega}_j \cup \bar{\Omega}'_j) \leq 1 - p + p/2$ where $\bar{\Omega}_j$ is the complementary of Ω_j, which implies $\mathbb{P}_{\boldsymbol{f}_j}(\Omega_j \cap \Omega'_j) \geq p/2$. Thus, for any estimator $\hat{\boldsymbol{\mu}}$ and M large enough,

$$\sup_{\boldsymbol{f} \in \{\boldsymbol{f}_1, ..., \boldsymbol{f}_M\}} \mathbb{P}_{\boldsymbol{f}}\left(|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 - \min_{k=1,...,M} |A_k\mathbf{y} - \boldsymbol{f}|_2^2 \geq c\sigma^2 \log(M)/2\right) \geq p/2 =: p^*.$$

□

Supplementary material

3.9 Proof of Proposition 3.4

Proof of Proposition 3.4. Let $a \in (0,1)$. By definition of $\hat{J}$, we have for all $k = 1, ..., M$,

$$\begin{aligned}|\hat{\boldsymbol{\mu}}_{\hat{\jmath}} - \boldsymbol{f}|_2^2 &\leq |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \Delta_{\hat{\jmath}k} - \frac{a}{2}|\hat{\boldsymbol{\mu}}_{\hat{\jmath}} - \hat{\boldsymbol{\mu}}_k|_2^2 + \frac{a}{2}|\hat{\boldsymbol{\mu}}_{\hat{\jmath}} - \hat{\boldsymbol{\mu}}_k|_2^2, \\ &\leq |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \frac{1}{a}\max_{j,k=1,...,M}\left(a\Delta_{jk} - \frac{1}{2}|a\hat{\boldsymbol{\mu}}_j - a\hat{\boldsymbol{\mu}}_k|_2^2\right) \\ &\qquad + a(|\hat{\boldsymbol{\mu}}_{\hat{\jmath}} - \boldsymbol{f}|_2^2 + |\boldsymbol{f} - \hat{\boldsymbol{\mu}}_k|_2^2).\end{aligned}$$

By rearranging, we have almost surely

$$\begin{aligned}|\hat{\boldsymbol{\mu}}_{\hat{\jmath}} - \boldsymbol{f}|_2^2 &\leq \frac{1+a}{1-a}\min_{k=1,...,M}|\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \frac{\Xi}{a(1-a)}, \\ &\text{where } \Xi := \max_{j,k=1,...,M}\left(2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}'_j - \hat{\boldsymbol{\mu}}'_k) - 2\sigma^2\mathrm{Tr}(A'_j - A'_k) - \frac{1}{2}|\hat{\boldsymbol{\mu}}'_j - \hat{\boldsymbol{\mu}}'_k|_2^2\right),\end{aligned}$$

and for all $j = 1, ..., M$, $\hat{\boldsymbol{\mu}}'_j := a\hat{\boldsymbol{\mu}}_j = A'_j\mathbf{y} + \boldsymbol{b}'_j$, $A'_j := aA_j$, $\boldsymbol{b}'_j := a\boldsymbol{b}_j$, and $|||A'_j|||_2 \leq 1$. By Proposition 3.12, as in the proof of Theorem 3.1, we have $\Xi \leq 30\sigma^2(x + 2\log M)$ with probability greater than $1 - \exp(-x)$.

Set $\varepsilon = 3a$ and choose the absolute constant $c > 0$ such that for all $\varepsilon < c$, $(1+a)/(1-a) \leq 1 + \varepsilon$ and $1/(1-a) \leq 2$. □

3.10 Smoothness adaptation

Proof of Proposition 3.9. Because the ellipsoids are ordered, if $\boldsymbol{f} \in \Theta(\beta, L)$ then

$$\mathbb{E}|\boldsymbol{f} - \tilde{\boldsymbol{\mu}}|_2^2 \leq \min_{j:\beta_j \leq \beta}\mathbb{E}|\boldsymbol{f} - A_{\beta_j}\mathbf{y}|_2^2 + 60\sigma^2\log M \leq \min_{j:\beta_j\leq\beta} C^* n^{\frac{1}{2\beta_j+1}}(1 + o(1)).$$

If $\beta \in [\beta_j, \beta_{j+1})$ for some j, then $\beta_{j+1} - \beta_j = \beta_j/(\log(n)\log\log n)$ and simple algebra yields

$$n^{\frac{1}{2\beta_j+1} - \frac{1}{2\beta+1}} \leq n^{\frac{2\beta_{j+1}-2\beta_j}{(2\beta+1)(2\beta_j+1)}} = n^{\frac{2\beta_j}{(2\beta+1)(2\beta_j+1)\log(n)\log\log n}} \leq n^{\frac{1}{(2\beta+1)\log(n)\log\log n}} \leq e^{\frac{1}{3\log\log n}},$$

where we used that $\beta \geq 1$ for the last inequality.

Now assume that $\beta \geq \beta_M$. Let $\varepsilon_n = 1/(\log(n)\log\log n)$, and

$$c = 120\log(1 + \varepsilon_3)/\varepsilon_3.$$

By definition of M,

$$\beta_M = e^{M \log\left(1+\frac{1}{\log(n)\log\log n}\right)} \geq e^{120\log\log(n)\frac{\log(1+\varepsilon_n)}{\varepsilon_n}} \geq e^{c\log\log(n)} = \log(n)^c,$$

since the function $t \to \log(1+t)/t$ is decreasing and $n \geq 3$. A numerical approximation gives $c \geq 1.01$. Thus,

$$n^{\frac{1}{2\beta_M+1}} n^{\frac{-1}{2\beta+1}} \leq n^{\frac{1}{2\beta_M+1}} \leq e^{\frac{\log n}{2\beta_M}} \leq e^{\frac{1}{2\log(n)^{c-1}}}.$$

In summary we have proved that $\min_{j:\beta_j\leq\beta} n^{\frac{1}{2\beta_j+1}} \leq n^{\frac{1}{2\beta+1}}(1+o(1))$, thus

$$\sup_{\boldsymbol{f}\in\Theta(\beta,L)} \mathbb{E}|\boldsymbol{f}-\tilde{\boldsymbol{\mu}}|_2^2 \leq C^* n^{\frac{1}{2\beta+1}}(1+o(1)).$$

□

3.11 Convex aggregation

Lemma 3.15 (Maurey argument)**.** *Let m and Λ_m^M be defined in (3.42) and (3.43). Let $Q(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\Sigma\boldsymbol{\theta} + \boldsymbol{v}^T\boldsymbol{\theta} + a$ for some semi-definite matrix Σ, $\boldsymbol{v} \in \mathbb{R}^M$ and $a \in \mathbf{R}$. Then*

$$\min_{\boldsymbol{\theta}\in\Lambda_m^M} Q(\boldsymbol{\theta}) \leq \min_{\boldsymbol{\theta}\in\Lambda^M} Q(\boldsymbol{\theta}) + \frac{4\max_{j=1,\dots,M}\Sigma_{jj}}{m}. \tag{3.61}$$

Proof of Lemma 3.15. Let $\boldsymbol{\theta}^* \in \Lambda^M \in \operatorname{argmin}_{\boldsymbol{\theta}\in\Lambda^M} Q(\boldsymbol{\theta})$. Let η be a random variable valued in $\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_M\}$ such that $\mathbb{P}(\eta=\boldsymbol{e}_j)=\theta_j^*$ for all $j=1,\dots,M$, and let $\eta_1,\dots,\eta_m$ be m i.i.d. copies of η. The random variable $\bar{\eta} = \frac{1}{m}\sum_{q=1}^m$ is valued in Λ_m^M and $\mathbb{E}\bar{\eta} = \boldsymbol{\theta}^*$. A bias variance decomposition and the independence of $\eta_1,\dots,\eta_m$ yield

$$\mathbb{E}Q(\bar{\eta}) = Q(\boldsymbol{\theta}^*) + \frac{\mathbb{E}[(\eta_1-\boldsymbol{\theta}^*)^T\Sigma(\eta_1-\boldsymbol{\theta}^*)]}{m}.$$

Using the triangle inequality, $\mathbb{E}[(\eta_1-\boldsymbol{\theta}^*)^T\Sigma(\eta_1-\boldsymbol{\theta}^*)] \leq 2(\boldsymbol{\theta}^*)^T\Sigma\boldsymbol{\theta}^* + 2\mathbb{E}[\eta_1^T\Sigma\eta_1] \leq 4\max_{j=1,\dots,M}\Sigma_{jj}$. Since $\bar{\eta}$ is valued in Λ_m^M, $\min_{\boldsymbol{\theta}\in\Lambda_m^M} Q(\boldsymbol{\theta}) \leq \mathbb{E}Q(\bar{\eta})$ and the proof is complete. □

Proof of (3.45) of Proposition 3.10. The condition on M, n implies that $m \geq 1$ where m is defined in (3.42). Let $C > 0$ be an absolute constant whose value may change from line to line. Applying Theorem 3.1 yields that on an event of probability greater than $1-2\exp(-x)$,

$$\frac{1}{n}|\hat{\boldsymbol{\mu}}_{\Lambda_m^M} - \boldsymbol{f}|_2^2 \leq \min_{\boldsymbol{\theta}\in\Lambda_m^M} \frac{1}{n}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 + \frac{C\sigma^2(\log(|\Lambda_m^M|)+x)}{n}. \tag{3.62}$$

By [65, page 8] we have

$$\log|\Lambda_m^M| \leq m\log\frac{2eM}{m}.$$

We use (3.61) with $Q(\boldsymbol{\theta}) = |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2$ to get

$$\min_{\boldsymbol{\theta}\in\Lambda_m^M} \frac{1}{n}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 \leq \min_{\boldsymbol{\theta}\in\Lambda^M} \frac{1}{n}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 + \frac{4}{nm}\max_{j=1,\dots,M}|\hat{\boldsymbol{\mu}}_j|_2^2.$$

We have $(1/n)\max_{j=1,\dots,M}|\hat{\boldsymbol{\mu}}_j|_2^2 \le C(|\boldsymbol{\xi}|_2^2/n + R^2) \le C(\sigma^2(2+3x)+R^2)$ on an event of probability at least $1-\exp(-x)$, where for the second inequality we used (3.21) with $B = I_{n\times n}$. Thus, with probability greater than $1-e^{-x}$,

$$\min_{\boldsymbol{\theta}\in\Lambda_m^M} \frac{1}{n}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 \le \min_{\boldsymbol{\theta}\in\Lambda^M} \frac{1}{n}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 + \frac{C(\sigma^2(2+3x)+R^2)}{m}. \tag{3.63}$$

Simple algebra yields that

$$\frac{1}{m} \le C\sqrt{\frac{\log(1+M/\sqrt{n})}{n}}, \qquad \frac{m\log(2eM/m)}{n} \le C\sqrt{\frac{\log(1+M/\sqrt{n})}{n}}. \tag{3.64}$$

Combining (3.62), (3.63) and (3.64) with the union bound completes the proof. □

Proof of (3.44) of Proposition 3.10. Let $\boldsymbol{\theta} \in \Lambda^M$. By definition of $\hat{\boldsymbol{\theta}}_C$, $C_p(\hat{\boldsymbol{\theta}}) \le C_p(\boldsymbol{\theta})$. This can be rewritten

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_C} - \boldsymbol{f}|_2^2 \le |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 + 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}).$$

The function $(\boldsymbol{\theta}', \boldsymbol{\theta}) \to 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})$ is bilinear, thus it is maximized at vertices, and

$$2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}) \le \max_{j,k=1,\dots,M} 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j) = \max_{j,k=1,\dots,M} \Delta_{jk},$$

where Δ_{jk} is defined in (3.19). Fix some pair (j,k). Let $B = A_j - A_k$ and $\boldsymbol{b} = (A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k$. We have $|||B|||_2 \le 2$, $\|B\|_{\mathrm{F}} \le |||B|||_2 \|I_{n\times n}\|_{\mathrm{F}} \le 2\sqrt{n}$ and $|\boldsymbol{b}|_2 \le 4R\sqrt{n}$. We apply (3.21) to the matrix B and (3.20) to the vector $\boldsymbol{b}$, which yields that with probability greater than $1-2\exp(-x)$,

$$\Delta_{jk} \le 8(\sigma^2 + \sigma R\sqrt{2})\sqrt{nx} + 8\sigma^2 x.$$

The union bound over all pairs $j,k = 1,\dots,M$ completes the proof. □

3.12 Sparsity oracle inequalities

3.12.1 Concentration inequalities for subgaussian vectors

A direct consequence of assumption (3.48) on the random vector $\boldsymbol{\xi}$ is the following Hoeffding-type concentration inequality:

$$\mathbb{P}\left(\alpha^T\boldsymbol{\xi} > K|\alpha|_2\sqrt{2x}\right) \le \exp(-x). \tag{3.65}$$

The following concentration inequality was proven in [55].

Proposition 3.16 (One sided concentration [55])**.** *Let $\boldsymbol{\xi}$ be a random vector in $\mathbb{R}^n$ satisfying (3.48) for some $K > 0$. Let A be a real $n\times n$ positive semi-definite symmetric matrix. Then for all $x > 0$, with probability greater than $1-\exp(-x)$,*

$$\boldsymbol{\xi}^T A\boldsymbol{\xi} \le K^2\left(\mathrm{Tr}A + 2\|A\|_{\mathrm{F}}\sqrt{x} + 2|||A|||_2 x\right). \tag{3.66}$$

This result is remarkable as it holds with the same constants as in the Gaussian case (3.21), under the weak assumption (3.48).

Corollary 3.17 (Corollary of Proposition 3.16 for any real matrix A)**.** *Under* (3.48) *and for any real matrix A, with probability greater than $1-\exp(-x)$, the following holds:*

$$\boldsymbol{\xi}^T A\boldsymbol{\xi} \leq K^2\left(\|A\|_1 + 2\|A\|_{\mathrm{F}}\sqrt{x} + 2|||A|||_2 x\right). \tag{3.67}$$

Proof. To see this, let $A_s := \frac{1}{2}(A+A^T)$ and let $|A_s| := \sqrt{A_s^2}$, the square root of the positive semi-definite symmetric matrix A_s^2. By definition of $|A_s|$ and the triangle inequality,

$$\begin{aligned} \boldsymbol{\xi}^T A\boldsymbol{\xi} = \boldsymbol{\xi}^T A_s\boldsymbol{\xi} \leq \boldsymbol{\xi}^T|A_s|\boldsymbol{\xi}, \qquad & \mathrm{Tr}(|A_s|) = \|A_s\|_1 \leq \|A\|_1, \\ |||\,|A_s|\,|||_2 = |||A_s|||_2 \leq |||A|||_2, \qquad & \|\,|A_s|\,\|_{\mathrm{F}} = \|A_s\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}}. \end{aligned}$$

Thus applying (3.66) to the matrix $|A_s|$ proves (3.67). □

In [55], the authors prove the following oracle inequality for the Least Squares estimator $\hat{\boldsymbol{\mu}}_V^{LS}$ on a d-dimensional linear subspace V of $\mathbb{R}^n$. The Least Squares estimator $\hat{\boldsymbol{\mu}}_V^{LS}$ is defined as the orthogonal projection of $\mathbf{y}$ on the linear subspace V.

Lemma 3.18 ([55])**.** *Under* (3.48)*, with probability greater than $1-\exp(-x)$:*

$$\begin{aligned} |\hat{\boldsymbol{\mu}}_V^{LS} - \boldsymbol{f}|_2^2 &\leq \min_{\mu\in V} |\mu - \boldsymbol{f}|_2^2 + K^2(d + 2\sqrt{dx} + 2x), \\ &\leq \min_{\mu\in V} |\mu - \boldsymbol{f}|_2^2 + K^2(2d + 3x). \end{aligned} \tag{3.68}$$

3.12.2 Preliminary result

Under the assumption (3.48), the authors of [55] proved the concentration inequality (3.66) and we use this concentration result to prove the following oracle inequality for aggregation of Least Squares estimators. Given an estimator $\hat{K}^2$, define for any $\boldsymbol{\theta} \in \Lambda^M$

$$U(\boldsymbol{\theta}) = |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 - 2\mathbf{y}^T\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + \tfrac{1}{2}\mathrm{pen}(\boldsymbol{\theta}) + 32\hat{K}^2\sum_{j=1}^M \theta_j \log\frac{1}{\pi_j},$$

where $\mathrm{pen}(\cdot)$ is the penalty (3.8). We consider the estimator $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}}$ of $\boldsymbol{f}$ where

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\Lambda^M} U(\boldsymbol{\theta}). \tag{3.69}$$

The function U is equal to the sum of H_{pen} (3.7) and some linear function of $\boldsymbol{\theta}$. Thus U is also convex.

Proposition 3.19. *Let $K>0$ be the smallest positive number such that the random vector $\boldsymbol{\xi}$ satisfies* (3.48)*. For all $j=1,\dots,M$, let $\mathbf{b}_j \in \mathbb{R}^n$ and let A_j be a square matrix of size n that satisfies $A_j = A_j^T = A_j^2$. Let $(\pi_1,\dots,\pi_M) \in \Lambda^M$ such that for all $j=1,\dots,M$, $\mathrm{Tr}(A_j) \leq \log(\pi_j^{-1})$. Let $\hat{K}>0$ be a given estimator and let $\hat{\boldsymbol{\theta}}$ be defined in (3.69). Let $\delta := \mathbb{P}(\hat{K}^2 < K^2)$. Then for all $x>0$, with probability greater than $1-\delta-2\exp(-x)$,*

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,\dots,M}\left(|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 64\hat{K}^2\log\frac{1}{\pi_j}\right) + 28K^2x.$$

Proposition 3.19 is proved below. Compared to (3.31), this oracle inequality holds for orthogonal projectors under the constraint $\mathrm{Tr}(A_j) \leq \log(\pi_j^{-1})$ for all $j = 1, ..., M$. However, this oracle inequality presents some advantages. First, it holds under (3.48) which is weaker that the Gaussian assumption of Theorem 3.6 since the noise coordinates do not need to be independent. Second, the quantity K^4/σ^2 that appears in (3.37) is not present here, which is possible thanks to the constraint $\mathrm{Tr}(A_j) \leq \log(\pi_j^{-1})$. This repairs the drawback of the right hand side of (3.37) which may be large if the noise random variables have a pathologically small variance compared to their subgaussian norm. Finally, one does not need to know the variance of the noise in order to compute the proposed estimator; its construction only relies on $\hat{K}$ which can be any estimate that *upper bounds* the subgaussian norm of the random vector $\boldsymbol{\xi}$. For instance, assume that $\boldsymbol{\xi}$ is zero-mean Gaussian with covariance matrix $\sigma^2 I_{n\times n}$, and assume that an estimator $\hat{\sigma}^2$ of σ^2 is accessible, and that this estimator has bounded bias. Let $\gamma > 1$ and $\varepsilon = \mathbb{P}(\hat{\sigma}^2 < \sigma^2/\gamma)$. The quantity ε is likely to be small if $\hat{\sigma}^2$ has a bounded bias and γ is large enough. Then one can use the upper bound $\hat{K}^2 = \gamma\hat{\sigma}^2$ in Proposition 3.19, which yields that with probability greater than $1 - 3\varepsilon$,

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} \left(|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 64\gamma\hat{\sigma}^2 \log\frac{1}{\pi_j} \right) + 28\sigma^2 \log(1/\varepsilon).$$

Thus, γ is used to perform a trade-off between the probability estimate and the remainder term of the oracle inequality. By using an upper bound for $\hat{K}^2$ in Proposition 3.19 the oracle inequality holds with slightly worse constants but with high probability.

Proof of Proposition 3.19. Let $\hat{\beta} = 32\hat{K}^2$, Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}$ for notational simplicity. As in the proof of (3.24) in Section 3.4 or the proof of Theorem 3.6, by convexity of U we have that for all $k = 1, ..., M$,

$$\begin{aligned} |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 \leq & \quad |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \nabla U(\hat{\boldsymbol{\theta}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}), \\ = & \quad 2\hat{\beta}\log\frac{1}{\pi_k} + \sum_{j=1}^{M} \hat{\theta}_j \zeta_{jk} \\ \leq & \quad 2\hat{\beta}\log\frac{1}{\pi_k} + \max_{j=1,...,M} \zeta_{jk}, \end{aligned}$$

where for all $j, k = 1, ..., M$,

$$\zeta_{jk} := \quad 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j) - \hat{\beta}\log\frac{1}{\pi_j\pi_k} - \tfrac{1}{2}|\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j|_2^2.$$

Let $B_{jk} = A_k - A_j$, and note that $|||B_{jk}|||_2 \leq 2$ because A_k and A_j are orthogonal projectors. Using $\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_j = B_{jk}\boldsymbol{\xi} + (B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j)$, we get

$$\zeta_{jk} = \quad 2\boldsymbol{\xi}^T(A_k - A_j)\boldsymbol{\xi} + \boldsymbol{\xi}^T\alpha_{jk} - \tfrac{1}{2}|B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2^2 - \tfrac{1}{2}|B_{jk}\boldsymbol{\xi}|_2^2 - \hat{\beta}\log\frac{1}{\pi_j\pi_k},$$

where $\alpha_{jk} := 2(I_{n\times n} - \frac{1}{2}B_{jk}^T)(B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j)$. The vector α_{jk} satisfies

$$|\alpha_{jk}|_2 \leq 2(1 + \tfrac{1}{2}|||B_{jk}|||_2)|B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2 \leq 4|B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2.$$

We also have $-|B_{jk}\boldsymbol{\xi}|_2^2 \leq 0$ almost surely.

Let $x > 0$. We now apply the concentration inequality (3.67) to the matrix $2B_{jk}$ and the Hoeffding-type inequality (3.65) to the vector α_{jk}. Using the union bound, the following holds with probability greater than $1 - 2\exp(-x)$:

$$\begin{aligned}\zeta_{jk} \leq \quad & K^2\left(2\left\|B_{jk}\right\|_1 + 4|||B_{jk}|||_2 x + 4\left\|B_{jk}\right\|_{\mathrm{F}}\sqrt{x}\right) \\ & + 2K(1 + \tfrac{1}{2}|||B_{jk}|||_2)|B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2\sqrt{2x} - \tfrac{1}{2}|B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2^2 \\ & - \hat{\beta}\log\frac{1}{\pi_j\pi_k}.\end{aligned}$$

We bound from above the first line of the RHS of the previous display. By the triangle inequality, and the assumption $\mathrm{Tr}(A_j) \leq \log(\pi_j^{-1})$, we have $\left\|B_{jk}\right\|_1 \leq \mathrm{Tr}(A_j + A_k) \leq \log((\pi_j\pi_k)^{-1})$. Using simple inequalities,

$$\left\|B_{jk}\right\|_{\mathrm{F}}\sqrt{x} \leq (\left\|A_j\right\|_{\mathrm{F}} + \left\|A_k\right\|_{\mathrm{F}})\sqrt{x} \leq (\left\|A_j\right\|_{\mathrm{F}}^2 + \left\|A_k\right\|_{\mathrm{F}}^2 + 2x)/2 \leq \frac{1}{2}\log\frac{1}{\pi_j\pi_k} + x.$$

Thus, $2\left\|B_{jk}\right\|_1 + 4|||B_{jk}|||_2 x + 4\left\|B_{jk}\right\|_{\mathrm{F}}\sqrt{x} \leq K^2(12x + 4\log\frac{1}{\pi_j\pi_k})$.

We now bound from above the second line. We apply the inequality $st \leq \frac{s^2+t^2}{2}$ with $t = |B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2$ and $s = 2K(1 + \frac{1}{2}|||B_{jk}|||_2)\sqrt{2x}$:

$$\begin{aligned}& 2K(1 + \tfrac{1}{2}|||B_{jk}|||_2)|B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2\sqrt{2x} - \tfrac{1}{2}|B_{jk}\boldsymbol{f} + \boldsymbol{b}_k - \boldsymbol{b}_j|_2^2 \\ & \qquad = st - \frac{t^2}{2} \leq \frac{s^2}{2} = 4K^2(1 + \tfrac{1}{2}|||B_{jk}|||_2)^2 x \leq 16K^2 x.\end{aligned}$$

For any $x' > 0$, let $x_{jk} = x' + \frac{1}{\pi_j\pi_k}$. By setting $x = x_{jk}$, the above displays yield the following bound on ζ_{jk}, with probability greater than $1 - 2\pi_j\pi_k\exp(-x')$:

$$\zeta_{jk} \leq 28K^2 x_{jk} - (\hat{\beta} - 4K^2)\log\frac{1}{\pi_j\pi_k} = 28K^2 x' - (\hat{\beta} - 32K^2)\log\frac{1}{\pi_j\pi_k}.$$

Using a union bound, we obtain that on an event of probability greater than $1 - \delta - 2\sum_{j=1}^M\sum_{k=1}^M \pi_j\pi_k\exp(-x') = 1 - \delta - 2\exp(-x')$, we have $\hat{\beta} \geq 32K^2$ and

$$\max_{j,k=1,\ldots,M}\zeta_{jk} \leq 28K^2 x'.$$

□

3.12.3 Sparsity pattern aggregation

We now combine (3.19) and (3.68) to prove Theorem 3.11.

Proof of Theorem 3.11. Given a subset $J \subset \{1, ..., p\}$, the projection matrix A_J satisfies $\mathrm{Tr}(A_J) \leq |J| \leq \log(\pi_J^{-1})$ since the normalizing constant of the weights $(\pi_j)_{J\subseteq\{1,\ldots,p\}}$ is greater than 1 [93, Section 5.2.1]. The estimator $\hat{\boldsymbol{\mu}}_J^{LS}$ satisfies the oracle inequality (3.68) with $d \leq |J|$, where $|J|$ denotes the cardinal of J and d is the dimension of the linear span of the columns whose indices are in J.

Let $\bar{\boldsymbol{\theta}} \in \mathbb{R}^p$ be a minimizer of the right hand side of (3.49) and let $\bar{J} \subset \{1, ..., p\}$ be the support of $\bar{\theta}$, hence $|\bar{\theta}|_0 = |\bar{J}|$. Since the RHS of (3.49) is random, $\bar{\theta}$ and its support are also random.

Let $t > 0$. For each support $J \subset \{1, ..., p\}$, the oracle inequality (3.68) applied to $x = t + \log(\pi_J^{-1})$ yields that with probability greater than $1 - \pi_J \exp(-t)$,

$$|\hat{\boldsymbol{\mu}}_J^{LS} - \boldsymbol{f}|_2^2 \leq |\mathbb{X}\bar{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 + K^2 \left(2|\bar{\theta}|_0 + 3\log\left(\frac{1}{\pi_J}\right) + 3t\right). \tag{3.70}$$

With the union bound, (3.70) holds simultaneously for all $J \subset \{1, ..., p\}$ with probability greater than $1 - \exp(-t) = 1 - \sum_{J \subset \{1,...,p\}} \pi_J \exp(-t)$.

We apply the oracle inequality of Proposition 3.19 and the oracle inequality (3.70) to $\hat{\boldsymbol{\mu}}_{\bar{J}}^{LS}$. With the union bound, we have with probability greater than $1-\delta-3\exp(-t)$:

$$|\mathbb{X}\hat{\theta}^{SPA} - \boldsymbol{f}|_2^2 \leq |\hat{\boldsymbol{\mu}}_{\bar{J}}^{LS} - \boldsymbol{f}|_2^2 + 64\hat{K}^2 \log\frac{1}{\pi_{\bar{J}}} + 28K^2 t,$$
$$|\hat{\boldsymbol{\mu}}_{\bar{J}}^{LS} - \boldsymbol{f}|_2^2 \leq |\mathbb{X}\bar{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 + K^2 \left(2|\bar{\theta}|_0 + 3\log\left(\frac{1}{\pi_{\bar{J}}}\right) + 3t\right),$$

where $A_{\bar{J}}$ is the projection matrix such that $\hat{\boldsymbol{\mu}}_{\bar{J}}^{LS} = A_{\bar{J}}\mathbf{y}$. We now use the following bound from [93, Section 5.2.1]:

$$\log\frac{1}{\pi_{\bar{J}}} \leq 2|\bar{\theta}|_0 \log\left(\frac{ep}{1 \vee |\bar{\theta}|_0}\right) + \frac{1}{2}.$$

Summing the two oracle inequalities above and applying the upper bound on $\log\frac{1}{\pi_{\bar{J}}}$ completes the proof. □

Chapter 4

Aggregation of supports along the Lasso path

4.1 Introduction

Let n, p be two positive integers. We consider the mean estimation problem

$$Y_i = \mu_i + \xi_i, \qquad i = 1, ..., n,$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathbb{R}^n$ is unknown, $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$ is a subgaussian vector, that is,

$$\mathbb{E}[\exp(\boldsymbol{v}^T \boldsymbol{\xi})] \leq \exp \frac{\sigma^2 |\boldsymbol{v}|_2^2}{2} \qquad \text{for all } \boldsymbol{v} \in \mathbb{R}^n, \tag{4.1}$$

where $\sigma > 0$ is the noise level and $|\cdot|_2$ is the Euclidean norm in $\mathbb{R}^n$. We only observe $\mathbf{y} = (Y_1, ..., Y_n)^T$ and wish to estimate $\boldsymbol{\mu}$. A design matrix $\mathbb{X}$ of size $n \times p$ is given and p may be larger than n. We do not require that the model is well-specified, i.e., that there exists $\boldsymbol{\beta}^* \in \mathbb{R}^p$ such that $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}^*$. Our goal is to find an estimator $\hat{\boldsymbol{\mu}}$ such that the prediction loss $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ is small, where $\|\cdot\|^2$ is the empirical loss defined by

$$\|\boldsymbol{u}\|^2 = \frac{1}{n}|\boldsymbol{u}|_2^2 = \frac{1}{n}\sum_{i=1}^n u_i^2, \qquad \boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n.$$

In a high-dimensional setting where $p > n$, the Lasso is known to achieve good prediction performance. For any tuning parameter $\lambda > 0$, define the Lasso estimate $\hat{\boldsymbol{\beta}}_\lambda^{\text{L}}$ as any solution of the convex minimization problem

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{L}} \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n}|\mathbf{y} - \mathbb{X}\boldsymbol{\beta}|_2^2 + \lambda |\boldsymbol{\beta}|_1 , \tag{4.2}$$

where $|\boldsymbol{\beta}|_1 = \sum_{j=1}^n |\beta_j|$ is the ℓ_1-norm. If $\mathbb{X}^T\mathbb{X}/n = I_{p\times p}$ where $I_{p\times p}$ is the identity matrix of size p, then an optimal choice of the tuning parameter is $\lambda_{univ} \sim \sigma\sqrt{\log(p)/n}$, up to a numerical constant. If the Restricted Eigenvalue condition holds (cf. Definition 4.1 below), then the universal tuning parameter $\lambda_{univ} \sim \sigma\sqrt{\log(p)/n}$ leads to good prediction performance [17]. However, if the columns of $\mathbb{X}$ are correlated and the Restricted Eigenvalue condition is not satisfied, the question of the optimal choice of the tuning parameter λ is still unanswered, even if the noise level σ^2 is known. Empirical and theoretical studies [100, 53, 38] have shown that if the columns of $\mathbb{X}$ are correlated, the Lasso estimate with a tuning parameter substantially smaller

than the universal parameter leads to a prediction performance which is substantially better than that of the Lasso estimate with the universal parameter. To summarize, these papers raise the following question:

Problem 4.1 (Data-driven selection of the tuning parameter)**.** *Find a data-driven quantity* $\hat{\lambda}$ *such that the prediction loss* $\|\boldsymbol{\mu} - \mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\hat{\lambda}}\|^2$ *is small with high probability.*

In this paper, we focus on a different problem, namely:

Problem 4.2 (Lasso Aggregation)**.** *Construct an estimator* $\hat{\boldsymbol{\mu}}$ *that mimics the prediction performance of the best Lasso estimator, that is, construct an estimator* $\hat{\boldsymbol{\mu}}$ *such that with high probability,*

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq C \min_{\lambda>0} \left(\|\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda} - \boldsymbol{\mu}\|^2 + \Delta(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda}) \right), \tag{4.3}$$

where $C \geq 1$ *is a constant and* $\Delta(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda})$ *is a small quantity.*

Problem 4.1 and Problem 4.2 have the same goal, that is, to achieve a small prediction loss with high probability. In Problem 4.1, the goal is to select a Lasso estimate that has small prediction loss. In Problem 4.2, we look for an estimator $\hat{\boldsymbol{\mu}}$ such that the prediction performance of $\hat{\boldsymbol{\mu}}$ is almost as good as the prediction performance of any Lasso estimate. The estimator $\hat{\boldsymbol{\mu}}$ may be of a different form than $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\hat{\lambda}}$ for some data-driven parameter $\hat{\lambda}$.

Our motivation to consider Problem 4.2 instead of Problem 4.1 is the following. Let $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_M$ be deterministic vectors $\mathbb{R}^n$. If the goal is to mimic the best approximation of $\boldsymbol{\mu}$ among $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_M$, it is well known in the literature on aggregation problems that an estimator of the form $\hat{\boldsymbol{\mu}} = \boldsymbol{f}_{\hat{k}}$ for some data-driven integer $\hat{k}$ is suboptimal (cf. Theorem 2.1 in [93], Section 2 of [58] and Proposition 6.1 in [46]). Thus, an optimal procedure cannot be valued in the discrete set $\{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_M\}$. Optimal procedures for this problem are valued in the convex hull of the set $\{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_M\}$. Examples are the Exponential Weights procedures proposed in [70, 35] or the Q-aggregation procedure of [33].

Although a lot of progress has been made for various aggregation problems, to our knowledge no previous work deals with the problem of aggregation of nonlinear estimators such as the collection $(\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda})_{\lambda>0}$ based on the sample. In the setting of the present paper, the observation $\mathbf{y}$ and the Lasso estimates are not independent: no data-split is performed and the same data is used to construct the Lasso estimators and to aggregate them.

We will show that aggregation of nonlinear estimators of the form $\mathbb{X}\hat{\boldsymbol{\beta}}$ is possible, for any nonlinear estimators $\hat{\boldsymbol{\beta}}$ and without any assumption on $\mathbb{X}$. For instance, an estimator $\hat{\boldsymbol{\mu}}$ that achieves (4.3) with

$$\Delta(\boldsymbol{\beta}) \simeq \frac{\sigma^2 |\boldsymbol{\beta}|_0}{n} \log\left(\frac{ep}{|\boldsymbol{\beta}|_0 \vee 1}\right)$$

is given in Section 4.3. Here, $|\boldsymbol{\beta}|_0$ denotes the number of nonzero coefficients of $\boldsymbol{\beta}$ and $a \vee b = \max(a, b)$.

Given a design matrix $\mathbb{X}$, we call *support* any subset T of $\{1, ..., p\}$. The cardinality of T is denoted by $|T|$ and for $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathrm{supp}(\boldsymbol{\beta})$ is the set of indices $k = 1, ..., p$ such that $\beta_k \neq 0$. Given a support T, we denote by Π_T the square matrix of size n which is the orthogonal projection on the linear span of the columns of $\mathbb{X}$ whose indices belong to T. Denote by $\mathcal{P}(\{1, ..., p\})$ the set of all subsets of $\{1, ..., p\}$. We will consider the following problem.

Problem 4.3 (Aggregation of a data-driven collection of supports)**.** *Let $\hat{F}$ be a data-driven collection of supports, that is, an estimator valued in $\mathcal{P}(\{1,...,p\})$. Construct an estimator $\hat{\boldsymbol{\mu}}$ such that with high probability,*

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{T \in \hat{F}} \Big(\|\Pi_T \boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \Delta(T) \Big), \tag{4.4}$$

where $\Delta(\cdot)$ is a function that takes small values.

The set $\hat{F}$ is a family of supports. Let us emphasize that both its cardinality and its elements can depend on the data $\mathbf{y}$. Note that for any support T, $\Pi_T \boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}^*_T$ where $\boldsymbol{\beta}^*_T$ minimizes $|\mathbb{X}\boldsymbol{\beta} - \boldsymbol{\mu}|_2^2$ subject to $\beta_k = 0$ for all $k \notin T$. In Section 4.3, we construct an estimator $\hat{\boldsymbol{\mu}}$ that satisfies (4.4) with $\Delta(T) \simeq \sigma^2 |T| \log(p/|T|)/n$ for all nonempty supports T. In the literature on aggregation problems, one is given a collection of estimators $\{\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M\}$ where $M \geq 1$ is a deterministic integer and the goal is to mimic the best estimator in this collection, cf. [96] and the references therein. A novelty of the present paper is to consider aggregation of a collection of estimators, where the cardinality of the collection depends on the data.

The main contributions of the present paper are the following.

- In Section 4.2, we propose an estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ that satisfies the oracle inequality (4.4) with $\Delta(T) \simeq \hat{\sigma}^2 |T| \log(p/|T|)/n$ for all nonempty supports T, where $\hat{\sigma}^2$ is an estimator of the noise level. This estimator solves Problem 4.3. We explain in Corollary 4.2 how Section 4.2 can be used to construct a procedure that aggregates nonlinear estimators of the form $\mathbb{X}\hat{\boldsymbol{\beta}}$.

- Section 4.3 is devoted to Problem 4.2. Using the result from Section 4.2, we construct an estimator $\hat{\boldsymbol{\mu}}$ that satisfies (4.3) with $\Delta(\boldsymbol{\beta}) \simeq \sigma^2 |\boldsymbol{\beta}|_0 \log(p/|\boldsymbol{\beta}|_0)$. The computational complexity of the procedure is the sum of the complexity of the regularization path of the Lasso and the complexity of a convex quadratic program.

The proofs can be found in the appendix.

4.2 Aggregation of a data-driven family of supports

Throughout this section, let $\hat{F}$ be a data-driven collection of supports and let $\hat{\sigma}^2 \geq 0$ be a real valued estimator. Let $\hat{M}$ be the cardinality of $\hat{F}$, and let $(\hat{T}_j)_{j=1,...,\hat{M}}$ be supports such that

$$\hat{F} = \{\hat{T}_1, ..., \hat{T}_{\hat{M}}\}. \tag{4.5}$$

For all supports $T \subset \{1, ..., p\}$, define the weights [93]

$$\pi_T := \left(H_p \binom{p}{|T|} e^{|T|} \right)^{-1}, \qquad H_p := \frac{e - e^{-p}}{e - 1}.$$

Note that by construction, the constant H_p is greater than 1 and $\sum_{T \in \mathcal{P}(\{1,...,p\})} \pi_T = 1$ where $\mathcal{P}(\{1, ..., p\})$ is the set of all subsets of $\{1, ..., p\}$. Given a support T, the Least Squares estimator on the linear span of the covariates indexed by T is $\Pi_T \mathbf{y}$.

We will consider two estimators of $\boldsymbol{\mu}$ based on $\hat{F}$ and $\hat{\sigma}^2$. The first estimator is defined as follows. Define the criterion

$$\mathrm{Crit}_{\hat{\sigma}^2}(T) = |\mathbf{y} - \Pi_T \mathbf{y}|_2^2 + 18\hat{\sigma}^2 \log \frac{1}{\pi_T}.$$

We have

$$|T| \leq \log \frac{1}{\pi_T} \leq \frac{1}{2} + 2|T| \log(ep/|T|) \tag{4.6}$$

for any support T. The lower bound is a direct consequence of $H_p > 1$ and the upper bound is proved in [93, (5.4)]. As (4.6) holds, the above criterion is of the same nature as C_p, AIC, BIC and their variants, cf. [18]. Define the estimator

$$\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y}) \qquad \text{where} \qquad \hat{T}_{\hat{F},\hat{\sigma}^2} \in \operatorname*{argmin}_{T \in \hat{F}} \mathrm{Crit}_{\hat{\sigma}^2}(T). \tag{4.7}$$

The estimator (4.7) is the orthogonal projection of $\mathbf{y}$ onto the linear span of the columns of $\mathbb{X}$ whose indices are in $\hat{T}_{\hat{F},\hat{\sigma}^2}$. If $\hat{F}$ is not data-dependent, the procedure $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$ is close to the one studied in [18].

We now define a second estimator valued in the convex hull of $(\Pi_T \mathbf{y})_{T \in \hat{F}}$. Let $\hat{M}$ be the cardinality of $\hat{F}$, and let $(\hat{T}_j)_{j=1,\dots,\hat{M}}$ be supports such that (4.5) holds. For any $j = 1, \dots, \hat{M}$, let $\hat{\boldsymbol{\mu}}_j = \Pi_{\hat{T}_j} \mathbf{y}$. Define a simplex in $\mathbb{R}^M$ as follows:

$$\Lambda^M = \Big\{ \boldsymbol{\theta} \in \mathbb{R}^M, \quad \sum_{j=1}^{M} \theta_j = 1, \quad \forall j = 1 \dots \hat{M}, \quad \theta_j \geq 0 \Big\}.$$

For any $\boldsymbol{\theta} \in \mathbb{R}^M$, define $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \sum_{j=1}^{\hat{M}} \theta_j \hat{\boldsymbol{\mu}}_j$. For all $\boldsymbol{\theta} \in \Lambda^M$, let

$$H_{\hat{F},\hat{\sigma}^2}(\boldsymbol{\theta}) \coloneqq |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \mathbf{y}|_2^2 + \frac{1}{2}\mathrm{pen}(\boldsymbol{\theta}) + 26\hat{\sigma}^2 \mathcal{K}\theta. \tag{4.8}$$

where

$$\mathrm{pen}(\boldsymbol{\theta}) \coloneqq \sum_{j=1}^{\hat{M}} \theta_j |\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2. \tag{4.9}$$

The penalty (4.9) is inspired by recent works on the Q-aggregation procedure [32], and it was used to derive sharp oracle inequalities for aggregation of linear estimators [33, 11] and density estimators [12]. The penalty pushes $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ towards the points $\{\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_{\hat{M}}\}$. Finally, the term $\mathcal{K}\theta$ is another penalty that pushes the coordinate θ_j to 0 if the size of the support $\hat{T}_j$ is large.

Define the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ as any minimizer of the function $H_{\hat{F},\hat{\sigma}^2}$ defined in (4.8):

$$\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2} \coloneqq \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}}, \qquad \hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Lambda^M} H_{\hat{F},\hat{\sigma}^2}(\boldsymbol{\theta}). \tag{4.10}$$

Theorem 4.1. *Let n, p be positive integers and let $\sigma > 0$. Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\mathbb{X}$ be any matrix of size $n \times p$. Let $\hat{F}$ be any data-driven collection of subsets of $\{1, \dots, p\}$. Assume that the noise $\boldsymbol{\xi}$ satisfies (4.1). Let $\hat{\sigma}^2$ be any real valued estimator and let $\delta \coloneqq \mathbb{P}(\hat{\sigma}^2 < \sigma^2)$. Then for all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ defined in (4.10) satisfies with probability greater than $1 - \delta - 2\exp(-x)$,*

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2} - \boldsymbol{\mu}\|^2 \leq \min_{T \in \hat{F}} \left(\|\Pi_T \boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \frac{\hat{\sigma}^2}{n} \left(24 + 96|T| \log\left(\frac{ep}{|T| \vee 1} \right) \right) \right) + \frac{22\sigma^2 x}{n}. \tag{4.11}$$

Furthermore, the estimator $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$ *satisfies with probability greater than* $1-\delta-2\exp(-x)$,

$$\|\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})-\boldsymbol{\mu}\|^2 \le \min_{T\in\hat{F}}\left(3\|\Pi_T\boldsymbol{\mu}-\boldsymbol{\mu}\|^2+\frac{\hat{\sigma}^2}{n}\left(26+104|T|\log\left(\frac{ep}{|T|\vee 1}\right)\right)\right)+\frac{28\sigma^2 x}{n}. \tag{4.12}$$

In previously studied aggregation problems, one is given a collection of estimators $\{\hat{\boldsymbol{\mu}}_1,...,\hat{\boldsymbol{\mu}}_M\}$ where $M\ge 1$ is a deterministic integer and the goal is to construct an estimator $\hat{\boldsymbol{\mu}}$ such that with high probability,

$$\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2 \le \min_{j=1,...,M}\|\hat{\boldsymbol{\mu}}_j-\boldsymbol{\mu}\|^2+\Delta_n(M),$$

where $\Delta_n(M)$ is a small error term that increases with M, cf. [96] and the references therein. Theorem 4.1 is of a different nature for several reasons. First, the set $\hat{F}$ is random, its cardinality can depend on the observed data $\mathbf{y}$. Second, the error term that appears inside the minimum of (4.11) does not depend on the cardinality of $\hat{F}$.

The estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ of Theorem 4.1 with $\hat{\sigma}^2=\sigma^2$ and $\hat{F}$ being the set of all subsets of $\{1,...,p\}$ was previously studied as the Exponential Screening estimator [92] or as the Sparsity Pattern Aggregate [93]. In this special case, $\hat{F}$ is deterministic and contains all the 2^p possible supports. Because of this exponential number of supports, computing the sparsity pattern aggregate in practice is hard. An MCMC algorithm is developed in [92] to compute an approximate solution of the sparsity pattern aggregate, but to our knowledge there is no theoretical guarantee that this MCMC algorithm will converge to a good approximation in polynomial time. The Sparsity Pattern Aggregate satisfies (4.11) with $\hat{\sigma}^2=\sigma^2$ and $\hat{F}=\mathcal{P}(\{1,...,p\})$. This sharp oracle inequality yields the minimax rate over all ℓ_q balls for all $0<q\le 1$, under no assumption on the design matrix $\mathbb{X}$ [33, 96].

To construct the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$, one has to solve the optimization problem (4.10). This is a convex quadratic program of size $|\hat{F}|$ with a simplex constraint. The complexity of computing $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ is polynomial in the cardinality of $\hat{F}$. Thus, if $\hat{F}$ is small then it is possible to construct $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ efficiently.

As the cardinality of $\hat{F}$ decreases, the prediction performance of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ becomes worse, but computing $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ becomes easier.

Problem 4.4. *Construct a data-driven set of supports* $\hat{F}$ *such that with high probability, there exists a support* $T\in\hat{F}$ *for which, simultaneously, the bias* $\|\Pi_T\boldsymbol{\mu}-\boldsymbol{\mu}\|^2$ *and the size* $|T|$ *are small.*

If we can construct such a set $\hat{F}$, by (4.11) the prediction loss of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ will be small. Note that Theorem 4.1 needs no assumption on the data-driven set $\hat{F}$ and the design matrix $\mathbb{X}$.

In the following Corollary, we perform aggregation of a family of nonlinear estimators of the form $(\mathbb{X}\hat{\boldsymbol{\beta}}_k)_{j\in J}$ for some set J. All estimators in the family share the same design matrix $\mathbb{X}$ and this matrix is deterministic.

Corollary 4.2. *Let* n,p *be positive integers and let* $\sigma>0$. *Let* $\boldsymbol{\mu}\in\mathbb{R}^n$ *and* $\mathbb{X}$ *be any matrix of size* $n\times p$. *Let* $\hat{F}$ *be any data-driven collection of subsets of* $\{1,...,p\}$. *Assume that the noise* $\boldsymbol{\xi}$ *satisfies* (4.1). *Let* $(\hat{\boldsymbol{\beta}}_j)_{j\in\hat{J}}$ *be a family of estimators valued in*

$\mathbb{R}^p$. *Both the cardinality of the family and its elements can depend on the data. Let* $\hat{\sigma}^2$ *be any real valued estimator and let* $\delta \coloneqq \mathbb{P}(\hat{\sigma}^2 < \sigma^2)$. *Define* $\hat{F} = \{\mathrm{supp}(\hat{\boldsymbol{\beta}}_j), j \in \hat{J}\}$ *and let* $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ *be the estimator* (4.10). *Then for all* $x > 0$, *the estimator* $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ *satisfies with probability greater than* $1 - \delta - 2\exp(-x)$,

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2} - \boldsymbol{\mu}\|^2 \leq \min_{j \in \hat{J}} \left(\|\mathbb{X}\hat{\boldsymbol{\beta}}_j - \boldsymbol{\mu}\|^2 + \frac{\hat{\sigma}^2}{n} \left(24 + 96 |\hat{\boldsymbol{\beta}}_j|_0 \log \left(\frac{ep}{|\hat{\boldsymbol{\beta}}_j|_0 \vee 1} \right) \right) \right) + \frac{22\sigma^2 x}{n}.$$

Using (4.12), a similar result can be readily obtained for the estimator $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$ with the leading constant 3.

4.3 Aggregation of supports along the Lasso path

Let us recall some properties of the Lasso path [44]. For a given observation $\mathbf{y}$, there exists a positive integer K and a finite sequence

$$\lambda_0 > \lambda_1 > \ldots > \lambda_K = 0$$

such that $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda = \mathbf{0}$ for all $\lambda > \lambda_0$, and such that

$$\forall \lambda \in (\lambda_{k+1}, \lambda_k), \qquad \mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda) = \mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_k}).$$

Thus, there is a finite number of supports on the Lasso path. In this section, we study the estimator of Theorem 4.1 in the special case $\hat{F} = \{\mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_k}), k = 0, \ldots, K\}$, that is, we aggregate all the supports that appear on the Lasso path.

Theorem 4.3. *Let* n, p *be positive integers and let* $\sigma > 0$. *Let* $\boldsymbol{\mu} \in \mathbb{R}^n$ *and* $\mathbb{X}$ *be any matrix of size* $n \times p$. *Assume that the noise* $\boldsymbol{\xi}$ *satisfies* (4.1). *Let* $\hat{\sigma}^2$ *be any real valued estimator and let* $\delta \coloneqq \mathbb{P}(\hat{\sigma}^2 < \sigma^2)$. *Let* $\lambda_0 > \ldots > \lambda_K$ *be the knots of the Lasso path. Let* $\hat{F} = \{\mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_j}), j = 0, \ldots, K\}$ *be the family of all supports that appear on the Lasso path and let* $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ *be the estimator* (4.10). *Then for all* $x > 0$, *the estimator* $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ *satisfies with probability greater than* $1 - \delta - 2\exp(-x)$,

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2} - \boldsymbol{\mu}\|^2 \leq \min_{\lambda > 0} \left(\|\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda - \boldsymbol{\mu}\|^2 + \frac{\hat{\sigma}^2}{n} \left(24 + 96 |\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda|_0 \log \left(\frac{ep}{|\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda|_0 \vee 1} \right) \right) \right) + \frac{22\sigma^2 x}{n}, \tag{4.13}$$

where for all $\lambda > 0$, $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda$ *is the Lasso estimator* (4.2).

Using (4.12), a similar result can be readily obtained for the estimator $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$ with the leading constant 3.

The computational complexity of the procedure of Theorem 4.3 is polynomial in the number of knots of the Lasso path. This will be further discussed in Section 4.4. In the rest of this section, we assume that $\hat{\sigma}^2 = \sigma^2$ and $\delta = 0$. We will come back to the estimation of the noise level in Section 4.5 below.

Interestingly, Theorem 4.3 does not need any assumption on the design matrix $\mathbb{X}$. The estimators $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ and $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$ have a good performance as soon as for some possibly unknown $\lambda > 0$, both the support of $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda$ and the loss $\|\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_\lambda - \boldsymbol{\mu}\|^2$ are small.

4.3.1 Prediction guarantees under the restricted eigenvalue condition

The goal of this section is to study the prediction performance of the procedure defined in Theorem 4.3 under the Restricted Eigenvalue condition on the design matrix $\mathbb{X}$.

Definition 4.1. For any $s \in \{1, ..., p\}$ and $c_0 > 0$, condition $RE(s, c_0)$ is satisfied if

$$\kappa(s, c_0) := \min_{T \subset \{1,...,p\}: |T| \leq s} \min_{\boldsymbol{\delta} \in \mathbb{R}^p : |\boldsymbol{\delta}_{T^c}|_1 \leq c_0 |\boldsymbol{\delta}_T|_1} \frac{|\mathbb{X}\boldsymbol{\delta}|_2}{\sqrt{n}|\boldsymbol{\delta}_T|_2} > 0.$$

The following result is a reformulation of Bickel et al. [17, Theorem 6.2].

Theorem 4.4 (Bickel et al. [17])**.** *Let $\mathbb{X}$ be such that the diagonal elements of $\mathbb{X}^T\mathbb{X}/n$ are all equal to 1. Assume that $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}^*$ and let $s := |\boldsymbol{\beta}^*|_0$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and that condition $RE(s, 3)$ is satisfied. Let $x_0 > 0$. There is an event $\Omega(x_0)$ of probability greater than $1 - e^{-x_0}$ on which the Lasso estimator (4.2) with tuning parameter $\lambda_{x_0} = \sigma\sqrt{8(x_0 + \log p)/n}$ satisfies simultaneously*

$$|\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_{x_0}}|_0 \leq \frac{64\phi_{max}}{\kappa^2(s, 3)} s, \tag{4.14}$$

$$\|\mathbb{X}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_{x_0}} - \boldsymbol{\beta}^*)\|^2 \leq \frac{128\sigma^2 s(x_0 + \log p)}{\kappa^2(s, 3)n}, \tag{4.15}$$

where ϕ_{max} is the largest eigenvalue of the matrix $\mathbb{X}^T\mathbb{X}/n$.

Thus, if the restricted eigenvalue condition is satisfied, the Lasso estimator with the universal parameter $\lambda_{x_0} = \sigma\sqrt{8(x_0 + \log p)/n}$ enjoys simultaneously an ℓ_0 norm of the same order as the true sparsity (cf. (4.14)), and a prediction loss of order $s \log(p)/n$ (cf. (4.15)).

Theorem 4.5 below is a direct consequence of Theorem 4.3 and the bounds (4.14)-(4.15).

Theorem 4.5. *Let n, p be positive integers and let $\sigma > 0$. Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\mathbb{X}$ be any matrix of size $n \times p$. Let $\hat{F}$ be any data-driven subset of $\{1, ..., p\}$. Assume that $\boldsymbol{\mu} = \mathbb{X}\boldsymbol{\beta}^*$ and let $s := |\boldsymbol{\beta}^*|_0$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and that condition $RE(s, 3)$ is satisfied.*

Let $\lambda_0 > ... > \lambda_K$ be the knots of the Lasso path. Let $\hat{F} = \{\mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_j}), j = 0, ..., K\}$ be the family of all supports that appear on the Lasso path and let $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2}$ be the estimator (4.10) with $\hat{\sigma}^2 = \sigma^2$. Then for all $x > 0$, the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2}$ satisfies with probability greater than $1 - 3\exp(-x)$,

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2} - \mathbb{X}\boldsymbol{\beta}^*\|^2 \leq \frac{(128 + 48\phi_{max})\sigma^2 s \log p}{\kappa^2(s, 3)n} + \frac{24\sigma^2}{n} + \frac{128\sigma^2 sx}{\kappa^2(s, 3)n} + \frac{22\sigma^2 x}{n}. \tag{4.16}$$

Furthermore,

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2} - \mathbb{X}\boldsymbol{\beta}^*\|^2 \leq \frac{(128 + 48\phi_{max})\sigma^2 s \log p}{\kappa^2(s, 3)n} + \frac{384\sigma^2 s}{\kappa^2(s, 3)n} + \frac{90\sigma^2}{n}. \tag{4.17}$$

Using (4.12), a similar result can be readily obtained for the estimator $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$ with different constants.

Proof of Theorem 4.5. By Theorem 4.3 with $\delta = 0$, there is an event $\Omega_{agg}(x)$ of probability greater than $1 - 2e^{-x}$ such that on $\Omega_{agg}(x)$ we have

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2} - \mathbb{X}\boldsymbol{\beta}^*\|^2 \le \|\mathbb{X}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_x} - \boldsymbol{\beta}^*)\|^2 + \frac{\sigma^2}{n}\left(24 + 96|\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_x}|_0 \log\left(\frac{ep}{|\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda}|_0 \vee 1}\right)\right) + \frac{22\sigma^2 x}{n}.$$

Let $\Omega(x)$ be the event defined in Theorem 4.4. Using the simple inequality $\log(p/(|\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda}|_0 \vee 1)) \le \log p$, and the bounds (4.14)-(4.15), we obtain that (4.16) holds on the event $\Omega_{agg}(x) \cap \Omega(x)$. By the union bound, the event $\Omega_{agg}(x) \cap \Omega(x)$ has probability greater than $1 - 3e^{-x}$. Finally, (4.17) is obtained from (4.16) by integration. □

The procedure studied in Theorem 4.5 aggregates the supports along the Lasso path using the procedure (4.10). A similar result holds for the estimator $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$ with a leading constant equal to 3. Theorem 4.5 has the following implications.

First, if $x > 0$ is fixed, the prediction performance (4.16) of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2}$ is similar to that of the Lasso with the universal tuning parameter λ_x, up to a multiplicative factor that only involves numerical constants and the quantity ϕ_{max}. As soon as ϕ_{max} (the operator norm of $\mathbb{X}^T\mathbb{X}/n$) is bounded from above by a constant, the estimator studied in Theorem 4.5 enjoys the best known prediction guarantees.

Second, Theorem 4.5 implies that the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2}$ satisfies the prediction bound (4.16) simultaneously for all confidence levels. That is, (4.16) holds for all $x > 0$ with probability greater than $1 - 3e^{-x}$, in contrast with the Lasso estimator with the universal parameter λ_{x_0} which depends on a fixed confidence level $1 - e^{-x_0}$. The Lasso estimator with the universal parameter λ_{x_0} satisfies the prediction bound (4.15) only for the confidence level $1 - e^{-x_0}$, but to our knowledge it is not known whether the Lasso estimator with the universal parameter λ_{x_0} satisfies a similar bound for different confidence levels than $1 - e^{-x_0}$. In this regard, the estimator studied in Theorem 4.5 provides a strict improvement compared to the Lasso with the universal parameter.

Third, the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2}$ of Theorem 4.5 satisfies the bound (4.17), that is, a prediction bound in expectation. Again, to our knowledge, it is not known whether the Lasso estimator with the universal parameter satisfies a similar bound in expectation.

Assuming that the bound (4.15) is tight and putting computational issues aside, the prediction performance of the procedure $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\sigma^2}$ of Theorem 4.5 is substantially better than the performance of the Lasso with the universal parameter, as soon as ϕ_{max} is bounded from above by a constant.

An upper bound similar to (4.14) is given in [16, Theorem 3 and Remark 3]. Namely, [16] prove that the square-root Lasso estimator with the universal tuning parameter $\hat{\boldsymbol{\beta}}$ satisfies $|\hat{\boldsymbol{\beta}}|_0 \le Cs$ with high probability, where s is the sparsity of the true parameter and C is a constant that depends on the sparse eigenvalues of the matrix $\mathbb{X}^T\mathbb{X}/n$, cf. [16, Condition P]. This upper bound can be used instead of (4.14) to prove results similar to (4.16) where ϕ_{max} is replaced by a smaller constant that depends on the sparse eigenvalues of $\mathbb{X}^T\mathbb{X}/n$.

4.4 Computational complexity of the Lasso path and $\hat{\boldsymbol{\mu}}^{\mathrm{q}}_{\hat{F},\hat{\sigma}^2}$

Computing the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ of Theorem 4.3 is done in two steps:

1. Compute the full Lasso path and let $\hat{F} = \{\mathrm{supp}(\lambda_0), ..., \mathrm{supp}(\lambda_K)\}$ be all the supports that appear on the Lasso path, where $\lambda_0, ..., \lambda_K$ are the knots of the Lasso path.
2. Compute $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ as a solution of the quadratic program (4.10), where $\hat{F}$ is defined by Step 1.

(We assume that the complexity of computing $\hat{\sigma}^2$ is negligible compared to the complexity of Step 1 and Step 2 above). The time complexity of Step 2 is the complexity of a convex quadratic program of size $|\hat{F}| \leq K$, where K is the number of knots on the Lasso path. Thus, the global cost of computing the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ of Theorem 4.3 is polynomial in K.

There exist efficient algorithms to compute the entire Lasso path [44]. However, [72] proved that for some values of $\mathbb{X}$ and $\mathbf{y}$, the regularization path of the Lasso contains more than $3^p/2$ knots. Hence, for some design matrix $\mathbb{X}$ and some observation $\mathbf{y}$, an exact computation of the full Lasso path is not realizable in polynomial time. In order to fix this computational issue, [72] propose an algorithm that computes an approximate regularization path for the Lasso. For some fixed $\varepsilon > 0$, this algorithm is guaranteed to terminate with less than $O(1/\sqrt{\varepsilon})$ knots and the points on the approximate path have a duality gap smaller than ε. This approximation algorithm can be used instead of computing the exact Lasso path. That is, one may compute the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ where $\hat{F}$ is the collection of supports that appear on the approximate path computed by the algorithm of [72].

Another solution to avoid computational issues is as follows. Let M be a positive integer. Instead of computing the Lasso path, one may consider a grid of tuning parameters $\lambda_1, ..., \lambda_M > 0$ and aggregate the supports of corresponding Lasso estimates $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_1}, ...\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_M}$. The advantage of this approach is twofold. First, for all $j = 1, ..., M$ the Lasso estimate $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_j}$ can be computed by standard convex optimization solvers. Second, the time complexity of the procedure is guaranteed to be polynomial in M and p. For any $x > 0$, by Corollary 4.2, this procedure satisfies, with probability greater than $1 - 3e^{-x}$

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2} - \boldsymbol{\mu}\|^2 \leq \min_{j=1,...,M} \left(\|\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_j} - \boldsymbol{\mu}\|^2 + \frac{\hat{\sigma}^2}{n} \left(24 + 96 |\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_j}|_0 \log \left(\frac{ep}{|\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_j}|_0 \vee 1} \right) \right) \right) + \frac{22\sigma^2 x}{n}.$$

This oracle inequality is not a strong as (4.13). However, if at least one of the Lasso estimates $\{\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda_j}, j = 1, ..., M\}$ enjoys a small prediction loss and a small ℓ_0 norm, then the prediction loss of $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ is also small.

4.5 A fully data-driven procedure using the Square-Root Lasso

This section proposes a fully data-driven procedure, based on the Square-Root Lasso. The choice of grid comes from the empirical and theoretical observations that for a

correlated design matrix, there exists a tuning parameter smaller than the universal parameter which enjoys better prediction performance than the universal parameter [100, 53, 38].

1. Let $\lambda_{\max} = 2\sqrt{\log(p/0.01)/n}$ be the universal parameter of the Square-Root Lasso [16] with confidence level 0.01.
2. Let $\lambda_{\min}$ be a conservatively small value of the tuning parameter.
3. Let M be an integer.
4. Consider the geometric grid $\{\lambda_1, ..., \lambda_M\}$ such that

$$\lambda_j = \lambda_{\min} \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{((j-1)/M-1)}, \qquad j = 1, ..., M.$$

5. Compute the Square-Root Lasso estimators $\hat{\boldsymbol{\beta}}^{\mathrm{SQ}}_{\lambda_1}, ...\hat{\boldsymbol{\beta}}^{\mathrm{SQ}}_{\lambda_M}$ with parameters $\lambda_1, ..., \lambda_M$ (it is possible to perform this computation simultaneously for all $\lambda_1, ..., \lambda_M$, cf. [85] and the references therein).
6. Let $\hat{F} = \{\mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathrm{SQ}}_{\lambda_j}), j = 1, ..., M\}$ be the supports of the computed Square-Root Lasso estimators.
7. Let $\hat{\sigma}^2$ be the variance estimated by the Square-Root Lasso with the universal parameter $\lambda_{\max}$.
8. For this choice of $\hat{\sigma}^2$ and $\hat{F}$, return the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ or the estimator $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$.

This estimator $\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2}$ returned by this procedure enjoys the theoretical guarantee

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2} - \boldsymbol{\mu}\|^2 \leq \min_{j=1,...,M} \left(\|\mathbb{X}\hat{\boldsymbol{\beta}}^{\mathrm{SQ}}_{\lambda_j} - \boldsymbol{\mu}\|^2 + \frac{\hat{\sigma}^2}{n} \left(24 + 96 |\hat{\boldsymbol{\beta}}^{\mathrm{SQ}}_{\lambda_j}|_0 \log \left(\frac{ep}{|\hat{\boldsymbol{\beta}}^{\mathrm{SQ}}_{\lambda_j}|_0 \vee 1} \right) \right) \right) + \frac{22\sigma^2 x}{n}$$

with probability greater than $1 - 3e^{-x}$. A similar guarantee with leading constant 3 can be obtained for the estimator $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y})$ using (4.12).

4.6 Concluding remarks

We have presented two procedures (4.7) and (4.10) that aggregates a data-driven collection of supports $\hat{F}$. These procedures satisfy the oracle inequalities given in Theorem 4.1 above, which is the main result of the paper. Sections 4.3 and 4.4 study the situation where $\hat{F}$ is the collection of supports that appear along the Lasso path. These procedures may be used for other data-driven collections $\hat{F}$ as well.

These procedures allow one to perform a trade-off between prediction performance and computational cost. If $\hat{F}$ contains all the 2^p supports, these procedures achieve optimal prediction guarantees with no assumption on the design matrix $\mathbb{X}$, but can not be realized in polynomial time. On the other hand, if the cardinality of $\hat{F}$ is small (say, polynomial in n and p), then it is possible to compute the estimators (4.7) and (4.10) in polynomial time. In view of (4.3), one should look for a data-driven set $\hat{F}$ with the following properties.

1. The set $\hat{F}$ is small so that the estimators (4.10) and (4.7) can be computed rapidly,

2. The set $\hat{F}$ contains a support T such that $|T|$ and $\|\pi_T \boldsymbol{\mu} - \boldsymbol{\mu}\|^2$ are simultaneously small, so that the procedures (4.10) and (4.7) enjoy good prediction performance.

A natural choice for $\hat{F}$ is the collection of supports that appear along the Lasso path. This choice of $\hat{F}$ was studied in Sections 4.3 and 4.4. Another natural choice is to aggregate the supports of several hard-thresholded Lasso estimators, since the hard-thresholded Lasso is sign-consistent under weak conditions on the design [78, Definition 5 and Corollary 2]. Further research will investigate other means to construct a data-driven collection $\hat{F}$ such that the above two properties are satisfied.

Acknoledgements

We would like to thank Alexandre Tsybakov for helpful comments during the writing of this manuscript.

Appendix

4.7 Proof of Theorem 4.1

For any matrix $A \in \mathbb{R}^{n\times n}$, define the operator norm of A and the Frobenius norm of A by

$$|||A|||_2 := \sup_{|\boldsymbol{u}|_2^2=1} |A\boldsymbol{u}|_2, \qquad \|A\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(A^T A)},$$

respectively.

Proof of (4.11). For all $S, T \subset \{1, ..., p\}$, define the event

$$\Omega_{S,T} = \left\{ Z(S,T) \le 4\sigma^2|S| + 22\sigma^2 \left(\log \frac{1}{\pi_S \pi_S} + x \right) \right\},$$

where

$$Z(S,T) = 2\boldsymbol{\xi}^T(\Pi_S \mathbf{y} - \Pi_T \boldsymbol{\mu}) - \frac{1}{2}|\Pi_S \mathbf{y} - \Pi_T \mathbf{y}|_2^2. \tag{4.18}$$

Define the event $\mathcal{V} := \{\hat{\sigma}^2 \ge \sigma^2\}$. On the event $\mathcal{A} := \mathcal{V} \cap (\cap_{S,T\subset\{1,...,p\}} \Omega_{S,T})$, we have simultaneously for all supports S, T

$$Z(S,T) - 26\hat{\sigma}^2 \log \frac{1}{\pi_S} - 22\sigma^2 \log \frac{1}{\pi_T} \le 22\sigma^2 x + 4\sigma^2|S| - 4\sigma^2 \log \frac{1}{\pi_S} \le 22\sigma^2 x$$

where we have used that $\log \frac{1}{\pi_S} \ge |S|$, cf. (4.6). By Lemma 4.6, on the event $\mathcal{A}$ we have

$$|\hat{\boldsymbol{\mu}}^{\mathrm{Q}}_{\hat{F},\hat{\sigma}^2} - \boldsymbol{\mu}|_2^2 \le \min_{T\in\hat{F}} \left(|\Pi_T \boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 + (26\hat{\sigma}^2 + 22\sigma^2) \log \frac{1}{\pi_T} \right) + 22\sigma^2 x.$$

To obtain (4.11), we use (4.6) and the fact that on the event $\mathcal{V}$, $26\hat{\sigma}^2 + 22\sigma^2 \le 48\hat{\sigma}^2$.

It remains to bound from below the probability of the event $\mathcal{A}$. Denote by $\mathcal{B}^c$ the complement of any event $\mathcal{B}$. We proceed with the union bound as follows,

$$\mathbb{P}(\mathcal{A}^c) \le \mathbb{P}(\mathcal{V}^c) + \sum_{S,T\subset\{1,...,p\}} \mathbb{P}(\Omega_{S,T}^c).$$

By definition, $\delta = \mathbb{P}(\mathcal{V}^c)$ and for any $S, T \subset \{1, ..., p\}$, Lemma 4.7 with $t = x + \log \frac{1}{\pi_S \pi_T}$ yields that $\mathbb{P}(\Omega_{S,T}^c) \le \pi_S \pi_T 2\exp(-x)$. As $\sum_{S,T\subset\{1,...,p\}} \pi_S \pi_T = (\sum_{S\subset\{1,...,p\}} \pi_S)^2 = 1$, we have established that

$$\mathbb{P}(\mathcal{A}^c) \le \delta + 2\exp(-x).$$

□

The proof of (4.12) is close to the argument used in [18], cf. [48, Section 2.3] for a recent reference on model selection. The novelty of the present paper is to consider a data-driven collection of estimators.

Proof of (4.12). Let $\hat{\Lambda} = 18\hat{\sigma}^2$ and let $\hat{T} = \hat{T}_{\hat{F},\hat{\sigma}^2}$ for notational simplicity. By definition of $\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y}) = \Pi_{\hat{T}}\mathbf{y}$, for all $T \in \hat{F}$ we have $\mathrm{Crit}_{\hat{\sigma}^2}(\hat{T}) \leq \mathrm{Crit}_{\hat{\sigma}^2}(T)$ which can be rewritten as

$$\begin{aligned}|\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y}) - \boldsymbol{\mu}|_2^2 + \hat{\Lambda}\log\frac{1}{\pi_{\hat{T}}} &\leq |\Pi_T\mathbf{y} - \boldsymbol{\mu}|_2^2 + \hat{\Lambda}\log\frac{1}{\pi_T} + 2\boldsymbol{\xi}^T(\Pi_{\hat{T}}\mathbf{y} - \Pi_T\mathbf{y}),\\ &\leq |\Pi_T\boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 + \hat{\Lambda}\log\frac{1}{\pi_T} + 2\boldsymbol{\xi}^T\Pi_{\hat{T}}\boldsymbol{\xi} + 2\boldsymbol{\xi}^T(\Pi_{\hat{T}}\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}) - |\Pi_T\boldsymbol{\xi}|_2^2.\end{aligned} \tag{4.19}$$

Define the event $\mathcal{V} := \{\hat{\sigma}^2 \geq \sigma^2\}$. For all $S, T \subset \{1, ..., p\}$, define

$$\begin{aligned}W(S) &= 2\boldsymbol{\xi}^T\Pi_S\boldsymbol{\xi} - 10\sigma^2\log\frac{1}{\pi_S},\\ W'(S,T) &= 2\boldsymbol{\xi}^T(\Pi_S\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}) - 8\sigma^2\log\frac{1}{\pi_S\pi_T} - \frac{1}{4}|\Pi_S\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}|_2^2.\end{aligned}$$

With this notation, using the simple inequality $-|\Pi_T\boldsymbol{\xi}|_2^2 \leq 0$, (4.19) implies that on the event $\mathcal{V}$,

$$|\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y}) - \boldsymbol{\mu}|_2^2 \leq |\Pi_T\boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 + \hat{\Lambda}\log\frac{1}{\pi_T} + 8\sigma^2\log\frac{1}{\pi_T} + W(\hat{T}) + W'(\hat{T},T) + \frac{1}{4}|\Pi_{\hat{T}}\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}|_2^2,$$

Using that $|\Pi_{\hat{T}}\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}|_2^2 \leq 2|\Pi_{\hat{T}}\boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 + 2|\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}|_2^2$ and that $|\Pi_{\hat{T}}\boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 \leq |\Pi_{\hat{T}}\mathbf{y} - \boldsymbol{\mu}|_2^2$, we obtain

$$\frac{1}{2}|\Pi_{\hat{T}_{\hat{F},\hat{\sigma}^2}}(\mathbf{y}) - \boldsymbol{\mu}|_2^2 \leq \frac{3}{2}|\Pi_T\boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 + \hat{\Lambda}\log\frac{1}{\pi_T} + 8\sigma^2\log\frac{1}{\pi_T} + W(\hat{T}) + W'(\hat{T},T).$$

For all $S, T \subset \{1, ..., p\}$, define the events

$$\Omega_S := \{W(S) \leq 6\sigma^2 x\}, \qquad \Omega_{S,T} := \{W'(S,T) \leq 8\sigma^2 x\}.$$

On the event $\mathcal{V} \cap (\cap_{S\subset\{1,...,p\}}\Omega_S) \cap (\cap_{S,T\subset\{1,...,p\}}\Omega_{S,T})$, (4.12) holds. It remains to bound from below the probability of this event.

For any fixed $S \subset \{1, ..., p\}$, using (4.6) and (4.23) with $t = x + \log\frac{1}{\pi_S}$ we have $\mathbb{P}(\Omega_S^c) \leq \pi_S e^{-x}$.

Let $S, T \subset \{1, ..., p\}$ be fixed. By using (4.22) with $\boldsymbol{v} = 2(\Pi_S\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu})$ and $t = x + \log\frac{1}{\pi_S\pi_T}$, we have that on an event of probability greater than $1 - \pi_S\pi_T e^{-x}$,

$$2\boldsymbol{\xi}^T(\Pi_S\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}) \leq 2\sigma\sqrt{2(x + \log(1/\pi_S\pi_T))}|\Pi_S\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}|_2 \leq 8\sigma^2\left(x + \log\frac{1}{\pi_S\pi_T}\right) + \frac{1}{4}|\Pi_S\boldsymbol{\mu} - \Pi_T\boldsymbol{\mu}|_2^2.$$

Thus, $\mathbb{P}(\Omega_{S,T}^c) \leq \pi_S\pi_T e^{-x}$.

As in the proof of (4.11), the union bound completes the proof. □

4.8 Technical Lemmas

Lemma 4.6. *For any estimator $\hat{\sigma}^2$, let $\hat{\boldsymbol{\theta}}$ be a minimizer of* (4.8). *Then, almost surely,*

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\mu}|_2^2 \leq \min_{q=1,...,\hat{M}}\left(|\Pi_{\hat{T}_q}\boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 + (26\hat{\sigma}^2 + 22\sigma^2)\log\frac{1}{\pi_{\hat{T}_q}}\right) + W, \tag{4.20}$$

where

$$W := \max_{S,T\in\hat{F}} \left(Z(S,T) - 26\hat{\sigma}^2 \log\frac{1}{\pi_S} - 22\sigma^2 \log\frac{1}{\pi_T} \right)$$

and $Z(\cdot,\cdot)$ *is defined in* (4.18).

Proof of Lemma 4.6. Let $\hat{\Lambda} = 26\hat{\sigma}^2$. The function $H_{\hat{F},\hat{\sigma}^2}$ is convex and differentiable, it can be rewritten as

$$\forall \boldsymbol{\theta} \in \Lambda^M,\ H_{\hat{F},\hat{\sigma}^2}(\boldsymbol{\theta}) = \frac{1}{2}|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 + |\mathbf{y}|_2^2 + \sum_{j=1}^{M} \theta_j \left(-2\mathbf{y}^T\hat{\boldsymbol{\mu}}_j + \frac{1}{2}|\hat{\boldsymbol{\mu}}_j|_2^2 + \hat{\Lambda}\log\frac{1}{\pi_{\hat{T}_j}} \right).$$

By simple algebra, for any $\boldsymbol{\theta'} \in \mathbb{R}^M$,

$$\nabla H_{\hat{F},\hat{\sigma}^2}(\hat{\boldsymbol{\theta}})^T\boldsymbol{\theta'} = \hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}}^T\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta'}} + \sum_{j=1}^{M} \theta'_j \left(-2\mathbf{y}^T\hat{\boldsymbol{\mu}}_j + \frac{1}{2}|\hat{\boldsymbol{\mu}}_j|_2^2 + \hat{\Lambda}\log\frac{1}{\pi_{\hat{T}_j}} \right), \tag{4.21}$$

$$\nabla H_{\hat{F},\hat{\sigma}^2}(\hat{\boldsymbol{\theta}})^T(-\hat{\boldsymbol{\theta}}) = -|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\mu}|_2^2 + |\boldsymbol{\mu}|_2^2 + \sum_{j=1}^{M} \hat{\boldsymbol{\theta}}_j \left(2\boldsymbol{\xi}^T\hat{\boldsymbol{\mu}}_j - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j|_2^2 - \hat{\Lambda}\log\frac{1}{\pi_{\hat{T}_j}} \right),$$

By summing the last display and equality (4.21) applied to $\boldsymbol{\theta'} = \boldsymbol{e}_q$, we get

$$\begin{aligned}\nabla H_{\hat{F},\hat{\sigma}^2}(\hat{\boldsymbol{\theta}})^T(\boldsymbol{e}_q - \hat{\boldsymbol{\theta}}) = -|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\mu}|_2^2 + |\hat{\boldsymbol{\mu}}_q - \boldsymbol{\mu}|_2^2 + \hat{\Lambda}\log\frac{1}{\pi_{\hat{T}_q}} \\ + \sum_{j=1}^{\hat{M}} \hat{\boldsymbol{\theta}}_j \left[2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_q) - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_q|_2^2 - \hat{\Lambda}\log\frac{1}{\pi_{\hat{T}_j}} \right].\end{aligned}$$

Since $\hat{\boldsymbol{\mu}}_q = \Pi_{\hat{T}_q}\mathbf{y}$ is a Least Squares estimator over the linear span of the covariates in $\hat{T}_q$, we have $|\hat{\boldsymbol{\mu}}_q - \mathbf{y}|_2^2 \leq |\Pi_{\hat{T}_q}\boldsymbol{\mu} - \mathbf{y}|_2^2$ which can be rewritten as

$$|\hat{\boldsymbol{\mu}}_q - \boldsymbol{\mu}|_2^2 \leq |\Pi_{\hat{T}_q}\boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 + 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}_q - \Pi_{\hat{T}_q}\boldsymbol{\mu}).$$

We thus have

$$\begin{aligned}\nabla H_{\hat{F},\hat{\sigma}^2}(\hat{\boldsymbol{\theta}})^T(\boldsymbol{e}_q - \hat{\boldsymbol{\theta}}) \leq -|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\mu}|_2^2 + |\Pi_{\hat{T}_q}\boldsymbol{\mu} - \boldsymbol{\mu}|_2^2 + (\hat{\Lambda} + 22\sigma^2)\log\frac{1}{\pi_{\hat{T}_q}} \\ + \sum_{j=1}^{\hat{M}} \hat{\boldsymbol{\theta}}_j \left[2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}_j - \Pi_{\hat{T}_q}\boldsymbol{\mu}) - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_q|_2^2 - \hat{\Lambda}\log\frac{1}{\pi_{\hat{T}_j}} - 22\sigma^2\log\frac{1}{\pi_{\hat{T}_q}} \right].\end{aligned}$$

For all $q = 1, ..., \hat{M}$, [21, Section 4.2.3] yields $\nabla H_{\hat{F},\hat{\sigma}^2}(\hat{\boldsymbol{\theta}})^T(\boldsymbol{e}_q - \hat{\boldsymbol{\theta}}) \geq 0$. Furthermore, a linear function over the simplex is maximized at a vertex, so almost surely we obtain (4.20). □

Lemma 4.7. *Let* $t > 0$. *For any supports* $S, T \subset \{1, ..., p\}$, *the quantity* $Z(S,T)$ *defined in* (4.18) *satisfies with probability greater than* $1 - 2\exp(-t)$,

$$Z(S,T) \leq 4\sigma^2|S| + 22\sigma^2 t.$$

Proof of Lemma 4.7. Let $D = \Pi_S - \Pi_T$. Then almost surely,

$$Z(S,T) = 2\boldsymbol{\xi}^T \Pi_S \boldsymbol{\xi} + \boldsymbol{\xi}^T (2D\boldsymbol{\mu} - D^2\boldsymbol{\mu}) - \frac{1}{2}|D\boldsymbol{\mu}|_2^2 - \frac{1}{2}|D\boldsymbol{\xi}|_2^2.$$

It is clear that $-|D\boldsymbol{\xi}|_2^2 \leq 0$. As $\boldsymbol{\xi}$ satisfies (4.1), a Chernoff bound yields that for all $\boldsymbol{v} \in \mathbb{R}^n$,

$$\mathbb{P}\left(\boldsymbol{\xi}^T \boldsymbol{v} > \sigma |\boldsymbol{v}|_2 \sqrt{2t}\right) \leq \exp(-t). \tag{4.22}$$

It is clear that $|||D|||_2 \leq 2$. We apply this concentration inequality to $\boldsymbol{v} = 2D\boldsymbol{\mu} - D^2\boldsymbol{\mu}$ to get that with probability greater than $1 - \exp(-t)$,

$$\begin{aligned}\boldsymbol{\xi}^T (2D\boldsymbol{\mu} - D^2\boldsymbol{\mu}) \leq \sigma |2D\boldsymbol{\mu} - D^2\boldsymbol{\mu}|_2 \sqrt{2t} &\leq \sigma |||2I_n - D|||_2 |D\boldsymbol{\mu}|_2 \sqrt{2t}, \\ &\leq \sigma 4 |D\boldsymbol{\mu}|_2 \sqrt{2t} \leq 16\sigma^2 t + \frac{1}{2}|D\boldsymbol{\mu}|_2^2.\end{aligned}$$

Finally, let $r \leq |S|$ be the rank of Π_S. The matrix Π_S is an orthogonal projector. Hence $\|\Pi_S\|_{\mathrm{F}}^2 = r$ and $|||\Pi_S|||_2 \leq 1$, so that applying the concentration inequality from [55] yields that with probability greater than $1 - \exp(-t)$,

$$2\boldsymbol{\xi}^T \Pi_S \boldsymbol{\xi} \leq 2\sigma^2 (r + 2\sqrt{rt} + 2t) \leq 4\sigma^2 r + 6\sigma^2 t \leq 4\sigma^2 |S| + 6\sigma^2 t. \tag{4.23}$$

A union bound completes the proof. □

Part II

From aggregation to shape restricted regression

Chapter 5

Sharp oracle bounds for monotone and convex regression through aggregation

Joint work with Alexandre Tsybakov.

We derive oracle inequalities for the problems of isotonic and convex regression using the combination of Q-aggregation procedure and sparsity pattern aggregation. This improves upon the previous results including the oracle inequalities for the constrained least squares estimator. One of the improvements is that our oracle inequalities are sharp, i.e., with leading constant 1. It allows us to obtain bounds for the minimax regret thus accounting for model misspecification, which was not possible based on the previous results. Another improvement is that we obtain oracle inequalities both with high probability and in expectation.

5.1 Introduction

Assume that we have the observations

$$Y_i = \mu_i + \xi_i, \qquad i = 1, ..., n, \tag{5.1}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathbb{R}^n$ is unknown, $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$ is a noise vector with n-dimensional Gaussian distribution $\mathcal{N}(0, \sigma^2 I_{n\times n})$ where $\sigma > 0$. We observe $\mathbf{y} = (Y_1, ..., Y_n)^T$ and we want to estimate $\boldsymbol{\mu}$. We can interpret μ_i as the values $f(X_i)$ of an unknown regression function $f : \mathcal{X} \to \mathbb{R}$ at given non-random points $X_i \in \mathcal{X}$, $i = 1, \dots, n$, where $\mathcal{X}$ is an abstract set. Then, the equivalent setting is that we observe $\mathbf{y}$ along with $(X_1, \dots, X_n)$ but the values of X_i are of no interest and can be replaced by their indices if we measure the loss in a discrete norm. Namely, for any $\boldsymbol{u} \in \mathbb{R}^n$ we consider the scaled (or the empirical) norm $\|\cdot\|$ defined by

$$\|\boldsymbol{u}\|^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2.$$

We will measure the error of an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ by the distance $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|$. Let $\mathcal{S}_n^{\uparrow}$ be the set of all non-decreasing sequences:

$$\mathcal{S}_n^{\uparrow} \coloneqq \{\boldsymbol{u} = (u_1, ..., u_n) \in \mathbb{R}^n : u_i \le u_{i+1}, \quad i = 1, ..., n-1\}.$$

For a subset $\mathcal{S}$ of $\mathcal{S}_n^\uparrow$, and any $\boldsymbol{\mu} \in \mathbb{R}^n$ the quantity $\min_{\boldsymbol{u}\in\mathcal{S}} \|\boldsymbol{u}-\boldsymbol{\mu}\|$ is the smallest approximation error achievable by a sequence in the set $\mathcal{S}$. This quantity defines a benchmark or oracle performance on $\mathcal{S}$. The accuracy of an estimator $\hat{\boldsymbol{\mu}}$ with respect to the oracle for any $\boldsymbol{\mu}$, not necessarily $\boldsymbol{\mu} \in \mathcal{S}$, can be characterized by the excess loss $\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\| - \min_{\boldsymbol{u}\in\mathcal{S}} \|\boldsymbol{u}-\boldsymbol{\mu}\|$. This is a measure of performance of $\boldsymbol{\mu}$ under model misspecification. One can also consider the expected quantities $R_1(\hat{\boldsymbol{\mu}},\boldsymbol{\mu}) = \mathbb{E}\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\| - \min_{\boldsymbol{u}\in\mathcal{S}} \|\boldsymbol{u}-\boldsymbol{\mu}\|$ or $R_2(\hat{\boldsymbol{\mu}},\boldsymbol{\mu}) = \mathbb{E}\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u}\in\mathcal{S}} \|\boldsymbol{u}-\boldsymbol{\mu}\|^2$ known under the name of regret measures. Here, $\mathbb{E}$ denotes the expectation with respect to the distribution of $\mathbf{y}$ satisfying (5.1). The minimax regret is defined as $\min_{\hat{\boldsymbol{\mu}}} \max_{\boldsymbol{\mu}\in\mathbb{R}^n} R_i(\hat{\boldsymbol{\mu}},\boldsymbol{\mu})$ for $i = 1, 2$, where $\min_{\hat{\boldsymbol{\mu}}}$ denotes the minimum over all estimators. We can characterize the performance of an estimator $\tilde{\boldsymbol{\mu}}$ by the closeness of its maximal regret $\max_{\boldsymbol{\mu}\in\mathbb{R}^n} R_i(\tilde{\boldsymbol{\mu}},\boldsymbol{\mu})$ to the minimax regret. This approach to measure the performance of estimators under model misspecification was pioneered by Vapnik and Chervonenkis who called it the criterion of minimax of the loss [101, Chapter 6]. In this paper, we follow this approach and establish non-asymptotic bounds for the maximal regret for some classes $\mathcal{S}$ of monotone and convex functions.

When the model is well-specified, i.e., the true function $\boldsymbol{\mu}$ belongs to the class $\mathcal{S}$, the approximation error vanishes and instead of the minimax regret it is natural to consider the minimax risk defined either as $\min_{\hat{\boldsymbol{\mu}}} \max_{\boldsymbol{\mu}\in\mathcal{S}} \mathbb{E}\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|$ or as $\min_{\hat{\boldsymbol{\mu}}} \max_{\boldsymbol{\mu}\in\mathcal{S}} \mathbb{E}\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2$ (the minimax squared risk). It is easy to see that the minimax risk is not greater than the minimax regret. A classical problem in nonparametric statistics is to study the behavior of minimax risks for different classes $\mathcal{S}$. In particular, there exist results concerning the minimax risks for classes of monotone and convex functions in our setting. We review some of them below. The behavior of the minimax regret is much less studied. For a recent overview and some general results we refer to [87] where it is shown that the rate of minimax regret can be different from that of the minimax risk. Note that [87] studies the prediction problem with i.i.d. observations, which is a setting different from ours.

A well-studied estimator under the monotonicity and convexity assumptions is the least squares estimator

$$\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}) \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathcal{S}} \|\mathbf{y}-\boldsymbol{u}\|^2.$$

In [82] it was shown that $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S})$ attains, up to logarithmic factors, the rates $n^{-2/3}$ and $n^{-4/5}$ of the mean squared risk for classes $\mathcal{S}$ of monotone and convex functions respectively and that these rates are optimal up to logarithmic factors when the minimax squared risk is used as a criterion. Under monotonicity constraints, the rate $n^{-2/3}$ was later observed in different settings, see for instance [7, 5].

One class of monotone functions we will be interested in here is defined as

$$\mathcal{S}_n^\uparrow(V) = \{\boldsymbol{\mu} \in \mathcal{S}_n^\uparrow : \ V(\boldsymbol{\mu}) \le V\}$$

where $V(\boldsymbol{\mu}) = \mu_n - \mu_1$ for any $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n) \in \mathcal{S}_n^\uparrow$, and $V > 0$ is a given constant. In [80, 107] it was shown that for any $\boldsymbol{\mu} \in \mathcal{S}_n^\uparrow$ we have

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2 \le c \max\left(\left(\frac{\sigma^2 V(\boldsymbol{\mu})}{n}\right)^{2/3}, \frac{\sigma^2 \log n}{n}\right) \tag{5.2}$$

for $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^\uparrow)$ and some absolute constant $c > 0$. This immediately implies an upper bound on the minimax risk on $\mathcal{S}_n^\uparrow(V)$. A recent paper [27] establishes the

oracle inequality

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \le C_* \min_{\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c_*\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right) \tag{5.3}$$

valid for all $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ where either $C_* = 6, c_* = 1$ [27, inequality (18)] or $C_* = 4, c_* = 4$ [27, inequality (30)]. Here, $k(\boldsymbol{u}) \ge 1$ for $\boldsymbol{u} = (u_1, \dots, u_n) \in \mathcal{S}_n^{\uparrow}$ is the integer such that $k(\boldsymbol{u}) - 1$ is the number of inequalities $u_i \le u_{i+1}$ that are strict for $i = 1, \dots, n-1$ (number of jumps of $\boldsymbol{u}$). Inequality (5.3) implies (up to a logarithmic factor) a bound as in (5.2) and also gives some more insight into the problem. For example, (5.3) shows that the fast rate $\frac{\log n}{n}$ is achieved if $\boldsymbol{\mu}$ has only one jump or a fixed, independent of n, number of jumps. This is not granted by (5.2).

Along with the least squares estimator, one may consider estimation of monotone functions via penalized least squares with total variation penalty. The corresponding estimator $\hat{\boldsymbol{\mu}}^{TV}$ is defined as

$$\hat{\boldsymbol{\mu}}^{TV} \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^n} \left(\frac{1}{2}\|\boldsymbol{u} - \mathbf{y}\|^2 + \lambda \sum_{i=1}^{n-1} |u_{i+1} - u_i| \right), \tag{5.4}$$

where $\lambda > 0$ is a tuning parameter. Statistical properties of this estimator were first studied in [74] where it was shown that $\|\hat{\boldsymbol{\mu}}^{TV} - \boldsymbol{\mu}\|$ attains the optimal rate $n^{-1/3}$ in probability on the class of functions of bounded variation (and thus on $\mathcal{S}_n^{\uparrow}(V)$). Recently, the performance of $\hat{\boldsymbol{\mu}}^{TV}$ was analyzed in [38] by considering $\hat{\boldsymbol{\mu}}^{TV}$ as a special instance of the Lasso estimator. If $\boldsymbol{\mu}^{\uparrow}$ is the projection of $\boldsymbol{\mu}$ onto $\mathcal{S}_n^{\uparrow}$, $\delta \in (0, 1)$ is a constant, and the tuning parameter λ is given by

$$\lambda = \sigma\sqrt{\frac{\log(n/\delta)}{k^* n}} \qquad \text{where } k^* = \left(\frac{V(\boldsymbol{\mu}^{\uparrow})^2 n \log(n/\delta)}{\sigma^2} \right)^{1/3},$$

the estimator $\hat{\boldsymbol{\mu}}^{TV}$ satisfies with probability greater than $1 - 2\delta$ the following oracle inequality [38, Proposition 6]:

$$\begin{aligned} \|\hat{\boldsymbol{\mu}}^{TV} - \boldsymbol{\mu}\|^2 \quad \le \quad & \|\boldsymbol{\mu}^{\uparrow} - \boldsymbol{\mu}\|^2 + 6 \left(\frac{\sigma^2 V(\boldsymbol{\mu}^{\uparrow}) \sqrt{\log(n/\delta)}}{n} \right)^{2/3} \\ & + \frac{2\sigma^2(1 + 2\log(1/\delta))}{n} \end{aligned} \tag{5.5}$$

for all $\boldsymbol{\mu} \in \mathbb{R}^n$. It follows from (5.5) that if the tuning parameter is chosen correctly, the estimator $\hat{\boldsymbol{\mu}}^{TV}$ achieves, up to a logarithmic factor, the minimax rate $n^{-2/3}$ in probability on the class $\mathcal{S}_n^{\uparrow}(V)$. Also, (5.5) implies a bound for the excess losses $\|\hat{\boldsymbol{\mu}}^{TV} - \boldsymbol{\mu}\|^i - \min_{\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}(V)} \|\boldsymbol{u} - \boldsymbol{\mu}\|^i$, $i = 1, 2$, corresponding to the class $\mathcal{S}_n^{\uparrow}(V)$. However, (5.5) does not allow us to evaluate the expected regrets $R_i(\hat{\boldsymbol{\mu}}^{TV}, \boldsymbol{\mu})$ since $\hat{\boldsymbol{\mu}}^{TV}$ depends on δ. It is also shown in [38, Proposition 4] that if $\lambda = 2\sigma\sqrt{(2/n)\log(n/\delta)}$, the estimator $\hat{\boldsymbol{\mu}}^{TV}$ satisfies

$$\|\hat{\boldsymbol{\mu}}^{TV} - \boldsymbol{\mu}\|^2 \le \min_{\boldsymbol{u}\in\mathbb{R}^n} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{4\sigma^2 k(\boldsymbol{u}) \log(n/\delta)}{n} r_n(\boldsymbol{u}) \right) \tag{5.6}$$

with probability greater than $1 - 2\delta$, where $k(\boldsymbol{u}) - 1$ for $\boldsymbol{u} \in \mathbb{R}^n$ is the number of jumps of $\boldsymbol{u}$, i.e., the cardinality of the set $\{i \in \{1, ..., n-1\} : u_i \ne u_{i+1}\}$,

$r_n(\boldsymbol{u}) = 3 + 256(\log(n) + (n/\Delta(\boldsymbol{u})))$ and $\Delta(\boldsymbol{u})$ is the minimum distance between two jumps in the sequence $\boldsymbol{u}$:

$$\Delta(\boldsymbol{u}) = \min\{d \geq 1 : \ \exists k \in \{1, ..., n\} \text{ with } u_{k+1} \neq u_k \text{ and } u_{k+d+1} \neq u_{k+d}\}.$$

The expressions on the right hand sides of (5.3) and (5.6) are small if the unknown sequence $\boldsymbol{\mu}$ is well approximated by a piecewise constant sequence with not too many pieces. In this regard, the two bounds have some similarity to sparsity oracle inequalities in high-dimensional linear regression, cf. [91, 93, 97]. This similarity can be easily explained as follows. Write (5.1) in the equivalent form

$$\mathbf{y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi},$$

with the matrix $\mathbb{X} = (X_{ij})_{i=1,...,n,\ j=1,...,n}$ where $X_{ij} = 1$ if $j \leq i$ and $X_{ij} = 0$ otherwise, and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_n^*)$ where $\beta_1^* = \mu_1$ and $\beta_i^* = \mu_i - \mu_{i-1}$ for $i = 2, \ldots, n$. With this notation, $k(\boldsymbol{\mu}) \in \{|\boldsymbol{\beta}^*|_0, 1 + |\boldsymbol{\beta}^*|_0\}$, where $|\boldsymbol{\beta}^*|_0$ denotes the number of non-zero components of $\boldsymbol{\beta}^*$. The value $k(\boldsymbol{\mu})$ is small when $\boldsymbol{\beta}^*$ is sparse. Thus, the problem of estimation of piecewise constant sequence $\boldsymbol{\mu}$ with small number of pieces can be considered as the problem of prediction in sparse linear regression with a specific design matrix $\mathbb{X}$. Similarly, we may write $\boldsymbol{u} = \mathbb{X}\boldsymbol{\beta}$, for $\boldsymbol{\beta}$ with components $\beta_1 = u_1$ and $\beta_i = u_i - u_{i-1}$ for $i = 2, \ldots, n$. These remarks suggest that we can apply the theory of sparsity oracle inequalities, in particular, sparsity pattern aggregation (cf. [91, 93, 97]) in the context of monotone estimation described above. Similar observation is valid for estimation under convexity constraints (see Section 5.3 below). In the present paper, we develop this argument using as a building block the Q-aggregation procedures [90, 32, 33, 11]. In particular, we construct an estimator $\hat{\boldsymbol{\mu}}$ such that

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right), \quad \forall\, \boldsymbol{\mu} \in \mathbb{R}^n, \tag{5.7}$$

for some absolute constant $c > 0$. Note that (5.7) is a sharp oracle inequality (i.e., an inequality with leading constant 1). It improves upon the oracle inequality (5.3) for the least squares estimator where the leading constant C_* is noticeably greater than 1 and the bound is valid only for $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$. The advantage of having leading constant 1 and arbitrary $\boldsymbol{\mu}$ in (5.7) is that it allows us to derive bounds on the excess risk and on the minimax regret, which was not possible based on the previous results. We also obtain sharp oracle inequalities with high probability for the same estimator. In addition, we show that it satisfies stronger sharp inequalities with the minimum $\min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}}$ on the right hand side of (5.7) replaced by $\min_{\boldsymbol{u} \in \mathbb{R}^n}$. This implies that our results are invariant to the direction of monotonicity; they remain valid if we replace everywhere monotone increasing by monotone decreasing functions. Finally, we derive similar results for the problem of estimation under the convexity constraints improving an oracle inequality obtained in [50].

5.2 Sparsity pattern aggregation for piecewise constant sequences

For any non-empty set $J \subseteq \{1, ..., n-1\}$, let $|J|$ denote the cardinality of J and define

$$\pi_J := \frac{\exp(-|J|)}{H \binom{n-1}{|J|}}, \qquad H := \sum_{i=0}^{n-1} \exp(-i). \tag{5.8}$$

Let $P_J \in \mathbb{R}^{n\times n}$ be the projector on the linear subspace V_J of $\mathbb{R}^n$ defined by

$$V_J := \Big\{ \boldsymbol{u} \in \mathbb{R}^n : \ \forall i \in \{1, ..., n-1\} \setminus J, \ u_{i+1} = u_i \Big\}.$$

In words, V_J is the space of all piecewise constant sequences that have jumps only at points in J. Given a vector $\mathbf{y}$ of observations and $\boldsymbol{\theta} = (\theta_J)_{J\subseteq\{1,...,n-1\}}$ where each $\theta_J \in \mathbb{R}$, let

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \sum_{J\subseteq\{1,...,n-1\}} \theta_J P_J \mathbf{y}.$$

Finally, let

$$\hat{\boldsymbol{\mu}}^Q = \boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}}$$

where $\hat{\boldsymbol{\theta}}$ is the solution of the optimization problem

$$\min_{\boldsymbol{\theta}\in\Lambda} \ \|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \mathbf{y}\|^2 + \sum_{J\subseteq\{1,...,n-1\}} \theta_J \left(\frac{2\sigma^2|J|}{n} + \frac{1}{2}\|\boldsymbol{\mu}_{\boldsymbol{\theta}} - P_J\mathbf{y}\|^2 + \frac{46\sigma^2}{n} \log\frac{1}{\pi_J} \right)$$

where

$$\Lambda = \left\{ \boldsymbol{\theta} : \ \theta_J \geq 0 \text{ for all } J \subseteq \{1, ..., n-1\}, \text{ and } \sum_{J\subseteq\{1,...,n-1\}} \theta_J = 1 \right\}.$$

This optimization problem is a convex quadratic program with a simplex constraint. It performs aggregation of the linear estimators $(P_J\mathbf{y})_{J\subseteq\{1,...,n-1\}}$ using the Q-aggregation procedure [32, 33, 11] with the prior weights (5.8). As the size of this quadratic program is of order 2^n, it is a computationally hard problem. The estimator $\hat{\boldsymbol{\mu}}^Q$ satisfies the following sharp oracle inequalities.

Theorem 5.1. *Let $\boldsymbol{\mu} \in \mathbb{R}^n$, $n \geq 2$, and assume that the noise vector $\boldsymbol{\xi}$ has distribution $\mathcal{N}(0, \sigma^2 I_{n\times n})$. There exist absolute constants $c, c' > 0$ such that for all $\delta \in (0, 1/3)$, the estimator $\hat{\boldsymbol{\mu}}^Q$ satisfies with probability at least $1 - 3\delta$,*

$$\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathbb{R}^n} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c\sigma^2 k(\boldsymbol{u}))}{n} \log\frac{en}{k(\boldsymbol{u})} \right) + \frac{c\sigma^2 \log(1/\delta)}{n}, \tag{5.9}$$

and

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathbb{R}^n} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c'\sigma^2 k(\boldsymbol{u})}{n} \log\frac{en}{k(\boldsymbol{u})} \right). \tag{5.10}$$

Proof. Let $J \subseteq \{1, ..., n-1\}$. Denote by $d = |J| + 1$ the dimension of the subspace V_J. Then, the projection estimator $P_J\mathbf{y}$ satisfies with probability at least $1 - \delta$ (see, for example, [55]):

$$\begin{aligned} \|P_J\mathbf{y} - \boldsymbol{\mu}\|^2 &\leq \|P_J\boldsymbol{\mu} - \boldsymbol{\mu}\|^2 + \frac{d + 2\sqrt{d\log(1/\delta)} + 2\log(1/\delta)}{n} \\ &\leq \min_{\boldsymbol{u}\in V_J} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{2(|J|+1) + 3\log(1/\delta)}{n}. \end{aligned} \tag{5.11}$$

The sharp oracle inequality from [11] yields that with probability at least $1 - 2\delta$ for all $J \subseteq \{1, ..., n-1\}$ we have

$$\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 \leq \|P_J\mathbf{y} - \boldsymbol{\mu}\|^2 + C\sigma^2 \log\frac{1}{\pi_J} + C\sigma^2\log(1/\delta), \tag{5.12}$$

for some absolute constant $C > 0$. Combining (5.11) and (5.12) with the union bound and the inequality (cf. [93, (5.4)]) $\log(1/\pi_J) \le 2(|J|+1)\log(en/(|J|+1)) + 1/2$, we find that with probability at least $1 - 3\delta$,

$$\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 \le \min_{J \subseteq \{1,\dots,n-1\}} \min_{\boldsymbol{u} \in V_J} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c\sigma^2(|J|+1)}{n} \log\left(\frac{en}{|J|+1}\right)\right) + c\sigma^2 \log(1/\delta)$$

where $c > 0$ is an absolute constant. Since $|J| + 1 = k(\boldsymbol{u})$ for all $\boldsymbol{u} \in V_J$ and $\min_{J \subseteq \{1,\dots,n-1\}} \min_{\boldsymbol{u} \in V_J} = \min_{\boldsymbol{u} \in \mathbb{R}^n}$, the bound (5.9) follows. Finally, (5.10) is obtained from (5.9) by integration. □

We now discuss some corollaries of Theorem 5.1. First, it follows that (5.7) is satisfied for $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^Q$, so the remarks after (5.7) apply. Next, in view of (5.10), for the class of monotone sequences with at most k jumps $\mathcal{S}_n^\uparrow(k) = \{\boldsymbol{u} \in \mathcal{S}_n^\uparrow : k(\boldsymbol{u}) \le k\}$ we have the following bounds for the maximal expected regrets

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\| - \min_{\boldsymbol{u} \in \mathcal{S}_n^\uparrow(k)} \|\boldsymbol{u} - \boldsymbol{\mu}\| \right) \le c\sqrt{\frac{\sigma^2 k}{n} \log\left(\frac{en}{k}\right)}, \tag{5.13}$$

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u} \in \mathcal{S}_n^\uparrow(k)} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \right) \le \frac{c\sigma^2 k}{n} \log\left(\frac{en}{k}\right), \tag{5.14}$$

where $c > 0$ is an absolute constant. The same bounds hold for the minimax risks over $\mathcal{S}_n^\uparrow(k)$ since the minimax risk is smaller than the minimax regret. Proposition 5.4 below shows that the bounds (5.13) and (5.14) are optimal up to logarithmic factors.

Finally, consider the consequences of Theorem 5.1 for the class $\mathcal{S}_n^\uparrow(V)$. To this end, define the integer k^* such that

$$k^* = \min\left\{ m \in \mathbb{N} : m \ge \left(\frac{V(\boldsymbol{\mu})^2 n}{\sigma^2 \log(en)} \right)^{1/3} \right\}$$

if the set $\left\{ m \in \mathbb{N} : m \ge \left(\frac{V(\boldsymbol{\mu})^2 n}{\sigma^2 \log(en)} \right)^{1/3} \right\}$ is non-empty, and $k^* = 1$ otherwise. We will need the following lemma.

Lemma 5.2. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^\uparrow$ and let $1 \le k \le n$ be an integer. Then there exists a sequence $\bar{\boldsymbol{u}} \in \mathcal{S}_n^\uparrow(k)$ such that*

$$\|\bar{\boldsymbol{u}} - \boldsymbol{\mu}\| \le \frac{V(\boldsymbol{\mu})}{2k}. \tag{5.15}$$

Next, there exists a sequence $\bar{\boldsymbol{u}} \in \mathcal{S}_{nk^*}^\uparrow$ such that

$$\|\bar{\boldsymbol{u}} - \boldsymbol{\mu}\|^2 \le \frac{1}{4} \max\left(\left(\frac{\sigma^2 V(\boldsymbol{\mu}) \log(en)}{n} \right)^{2/3}, \frac{\sigma^2 \log(en)}{n} \right). \tag{5.16}$$

In addition,

$$\frac{\sigma^2 k^*}{n} \log \frac{en}{k^*} \le 2 \max\left(\left(\frac{\sigma^2 V(\boldsymbol{\mu}) \log(en)}{n} \right)^{2/3}, \frac{\sigma^2 \log(en)}{n} \right). \tag{5.17}$$

Proof. To construct the sequence $\bar{\boldsymbol{u}}$, consider the k intervals

$$I_j = \left[\mu_1 + \frac{j-1}{k}V(\boldsymbol{\mu}), \mu_1 + \frac{j}{k}V(\boldsymbol{\mu})\right), \qquad j = 1, ..., k-1,$$

and $I_k = [\mu_1 + \frac{k-1}{k}V(\boldsymbol{\mu}), \mu_n]$. For all $j = 1, ..., k$, let

$$J_j = \{i = 1, ..., n : \quad \mu_i \in I_j\}.$$

For any $i \in \{1, ..., n\}$ there exists a unique $j \in \{1, ..., k\}$ such that $i \in I_j$. Let $\bar{u}_i = \mu_1 + \frac{j-1/2}{k}V(\boldsymbol{\mu})$ for all $i \in I_j$. Then the sequence $\bar{\boldsymbol{u}} = (\bar{u}_1, \ldots, \bar{u}_n)$ is non-decreasing, it has at most k pieces, i.e., $k(\bar{\boldsymbol{u}}) \le k$, and $|\bar{u}_i - \mu_i| \le \frac{V(\boldsymbol{\mu})}{2k}$ for $i = 1, ..., n$. Thus (5.15) follows. Next, note that if $k^* = 1$, then $V(\boldsymbol{\mu})^2 \le \sigma^2 \log(en)/n$. If $k^* > 1$, then by definition of k^*, $V(\boldsymbol{\mu})^2/(k^*)^2 \le (\sigma^2 V(\boldsymbol{\mu}) \log(en)/n)^{2/3}$. Thus, (5.16) follows. The bound (5.17) is straightforward by studying the cases $k^* = 1$ and $k^* > 1$ separately. □

We can now derive the following corollary of Theorem 5.1.

Corollary 5.3. *Under the assumptions of Theorem 5.1, there exists an absolute constant $c > 0$ such that, for any $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$,*

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 \le c \max\left(\left(\frac{\sigma^2 V(\boldsymbol{\mu}) \log n}{n}\right)^{2/3}, \frac{\sigma^2 \log n}{n}\right). \tag{5.18}$$

In addition, for any $V > 0$ and any $\boldsymbol{\mu} \in \mathbb{R}^n$ the expected regret of $\hat{\boldsymbol{\mu}}^Q$ satisfies

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\| - \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}(V)} \|\boldsymbol{u} - \boldsymbol{\mu}\| \ \le c \max\left(\left(\frac{\sigma^2 V \log n}{n}\right)^{1/3}, \sigma\sqrt{\frac{\log n}{n}}\right) \tag{5.19}$$

where $c > 0$ is an absolute constant.

Proof. Inequality (5.18) is straightforward in view of (5.10), (5.16), and (5.17). To prove (5.19), fix any $\boldsymbol{\mu} \in \mathbb{R}^n$ and consider

$$\boldsymbol{\mu}^* \in \operatorname*{argmin}_{\boldsymbol{\mu}' \in \mathcal{S}_n^{\uparrow}(V)} \|\boldsymbol{\mu}' - \boldsymbol{\mu}\|.$$

From (5.10) and the fact that the function $x \mapsto x \log\left(\frac{en}{x}\right)$ is increasing for $1 \le x \le n$ we get

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\| &\le \min_{\boldsymbol{u} \in \mathcal{S}_{nk^*}^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\| + \sqrt{c' \frac{\sigma^2 k^*}{n} \log\left(\frac{en}{k^*}\right)}\right) \\
&\le \min_{\boldsymbol{u} \in \mathcal{S}_{nk^*}^{\uparrow}} \|\boldsymbol{u} - \boldsymbol{\mu}^*\| + \|\boldsymbol{\mu}^* - \boldsymbol{\mu}\| + \sqrt{c' \frac{\sigma^2 k^*}{n} \log\left(\frac{en}{k^*}\right)} \\
&\le \|\boldsymbol{\mu}^* - \boldsymbol{\mu}\| + c'' \max\left(\left(\frac{\sigma^2 V \log n}{n}\right)^{1/3}, \sigma\sqrt{\frac{\log n}{n}}\right)
\end{aligned}$$

for an absolute constant $c'' > 0$ where the last inequality follows from (5.16) and (5.17). □

The estimator $\hat{\boldsymbol{\mu}}^Q$ shown in Theorem 5.1 satisfies the sharp oracle inequalities both in expectation and with high probability. Previous results for the least squares estimator [27] were only obtained in expectation and the results on the ℓ_1-penalized estimator (5.4) are only obtained with high probability.

Finally, the following result shows that the upper bounds (5.13) and (5.14) are optimal up to logarithmic factors.

Proposition 5.4. *Let $n \geq 2, V > 0$ and $\sigma > 0$. There exist absolute constants $c, c' > 0$ such that for any positive integer $k \leq n$ satisfying $k^3 \leq 16nV^2/\sigma^2$ we have*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}(k) \cap \mathcal{S}_n^{\uparrow}(V)} \mathbb{P}_{\boldsymbol{\mu}} \left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq \frac{c\sigma^2 k}{n} \right) > c',$$

where $\mathbb{P}_{\boldsymbol{\mu}}$ denotes the distribution of $\mathbf{y}$ satisfying (5.1) and $\inf_{\hat{\boldsymbol{\mu}}}$ is the infimum over all estimators.

For $k = 1, ..., n$, take any $V > 0$ large enough to satisfy $k^3 \leq 16nV^2/\sigma^2$. Then, Proposition 5.4 and Markov's inequality yield the following lower bounds on the minimax risks over the class $\mathcal{S}_n^{\uparrow}(k)$:

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}(k)} \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \geq c\sqrt{\frac{c'\sigma^2 k}{n}}, \quad \inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}(k)} \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq \frac{cc'\sigma^2 k}{n}. \tag{5.20}$$

As the minimax risk is smaller than the minimax regret, (5.20) also provides lower bounds for the corresponding minimax regrets over $\mathcal{S}_n^{\uparrow}(k)$. Combining this with (5.13) and (5.14) we find that the estimator $\hat{\boldsymbol{\mu}}^Q$ achieves up to logarithmic factors the optimal rate with respect to the minimax regret.

Next, Proposition 5.4 implies the following lower bound on the minimax deviation risk on $\mathcal{S}_n^{\uparrow}(V)$.

Corollary 5.5. *Let $n \geq 2, V > 0$ and $\sigma > 0$. There exist absolute constants $c, c' > 0$ such that*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}(V)} \mathbb{P}_{\boldsymbol{\mu}} \left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq c \max \left\{ \left(\frac{\sigma^2 V}{n} \right)^{2/3}, \frac{\sigma^2}{n} \right\} \right) > c'. \tag{5.21}$$

To prove this corollary it is enough to note that if $16nV^2/\sigma^2 \geq 1$, by choosing k in Proposition 5.4 as the integer part of $(16nV^2/\sigma^2)^{1/3}$, we obtain the lower bound corresponding to $\left(\frac{\sigma^2 V}{n}\right)^{2/3}$ under the maximum in (5.21). On the other hand, if $16nV^2/\sigma^2 < 1$ the term $\frac{\sigma^2}{n}$ is dominant, so that we need to have the lower bound of the order $\frac{\sigma^2}{n}$, which is trivial (it follows from a reduction to the bound for the class composed of two constant functions).

It follows from (5.21) and (5.18) that the estimator $\hat{\boldsymbol{\mu}}^Q$ achieves, up to logarithmic factors, the optimal rate with respect to the minimax risk on the class $\mathcal{S}_n^{\uparrow}(V)$. Using (5.19) and the fact that the minimax risk is smaller than the minimax regret, we conclude that it is also the optimal rate up to logarithmic factors for the minimax regret.

Proof of Proposition 5.4. We assume for simplicity that n is a multiple of k. The general case is treated analogously. For any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \{0,1\}^k$, let $d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') = |\{i =$

$1, ..., k : \omega_i \neq \omega'_i\}|$ be the Hamming distance between $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$. By the Varshamov-Gilbert bound [99, Lemma 2.9], there exists a set $\Omega \subset \{0,1\}^k$ such that

$$\mathbf{0} = (0, ..., 0) \in \Omega, \quad \log(|\Omega| - 1) \geq k/8, \quad \text{and} \quad d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') > k/8 \tag{5.22}$$

for any two distinct $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$. For each $\boldsymbol{\omega} \in \Omega$, define a vector $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathbb{R}^n$ with components

$$u_i^{\boldsymbol{\omega}} = \frac{\lfloor (i-1)k/n \rfloor \, V}{2k} + \gamma \omega_{\lfloor (i-1)k/n \rfloor + 1}, \qquad i = 1, ..., n,$$

where $\gamma = (1/8)\sqrt{\sigma^2 k/n}$, and $\lfloor x \rfloor$ denotes the maximal integer smaller than x. For any $\boldsymbol{\omega} \in \Omega$, $\boldsymbol{u}^{\boldsymbol{\omega}}$ is a piecewise constant sequence with $k(\boldsymbol{u}^{\boldsymbol{\omega}}) \leq k$, $\boldsymbol{u}^{\boldsymbol{\omega}}$ is a non-decreasing sequence because $\gamma \leq V/(2k)$, and by construction $V(\boldsymbol{u}^{\boldsymbol{\omega}}) \leq V$. Thus, $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathcal{S}_n^{\uparrow}(k) \cap \mathcal{S}_n^{\uparrow}(V)$ for all $\boldsymbol{\omega} \in \Omega$. Moreover, for any $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$,

$$\|\boldsymbol{u}^{\omega} - \boldsymbol{u}^{\boldsymbol{\omega}'}\|^2 = \frac{\gamma^2}{k} d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') \geq \frac{\gamma^2}{8} = \frac{\sigma^2 k}{512 n}.$$

Set for brevity $P_{\boldsymbol{\omega}} = \mathbb{P}_{\boldsymbol{u}^{\omega}}$. The Kullback-Leibler divergence $K(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}'})$ between $P_{\boldsymbol{\omega}}$ and $P_{\boldsymbol{\omega}'}$ is equal to $\frac{n}{2\sigma^2}\|\boldsymbol{u}^{\omega} - \boldsymbol{u}^{\boldsymbol{\omega}'}\|^2$ for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$. Thus,

$$K(P_{\boldsymbol{\omega}}, P_{\mathbf{0}}) = \frac{\gamma^2 n d_H(\mathbf{0}, \boldsymbol{\omega})}{2k\sigma^2} \leq \frac{k}{128} \leq \frac{\log(|\Omega| - 1)}{16}. \tag{5.23}$$

Applying [99, Theorem 2.7] with $\alpha = 1/16$ completes the proof. □

5.3 Estimation of convex sequences by aggregation

Assume that $n \geq 3$ and define the set of convex sequences $\mathcal{S}_n^{\cup}$ as follows:

$$\mathcal{S}_n^{\cup} = \{\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n : \; 2u_i \leq u_{i+1} + u_{i-1}, \; i = 2, \ldots, n-1\}.$$

For any $\boldsymbol{u} \in \mathbb{R}^n$, we introduce the integer $q(\boldsymbol{u}) \geq 1$ such that $q(\boldsymbol{u}) - 1$ is the cardinality of the set $\{i = 1, ..., n-1 : 2u_i \neq u_{i+1} + u_{i-1}\}$. If $\boldsymbol{u} \in \mathcal{S}_n^{\cup}$, $q(\boldsymbol{u}) - 1$ is the number of inequalities $2u_i \leq u_{i+1} + u_{i-1}$ that are strict for $i = 2, ..., n-1$. The value $q(\boldsymbol{u})$ is small if $\boldsymbol{u}$ is a piecewise linear sequence with a small number of pieces.

The performance of the least squares estimator over convex sequences $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\cup})$ has been recently studied in [50]. If the unknown vector $\boldsymbol{\mu}$ belongs to the set $\mathcal{S}_n^{\cup}$, [50] shows that the estimator $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\cup})$ satisfies the risk bound

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\cup}) - \boldsymbol{\mu}\|^2 \leq c \log(en)^{5/4} \left(\frac{\sigma^2 \sqrt{R(\boldsymbol{\mu})}}{n} \right)^{4/5},$$

where $R(\boldsymbol{\mu}) = \max(1, \min\{\|\boldsymbol{\tau} - \boldsymbol{\mu}\|^2, \boldsymbol{\tau} \text{ is affine}\})$ and $c > 0$ is an absolute constant. It is proved in [27, Example 2.3] that the least squares estimator $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\cup})$ satisfies the oracle inequality

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\cup}) - \boldsymbol{\mu}\|^2 \leq 6 \min_{\boldsymbol{u} \in \mathcal{S}_n^{\cup}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{c\sigma^2 q(\boldsymbol{u}) \log\left(\frac{en}{q(\boldsymbol{u})}\right)^{5/4}}{n} \right), \tag{5.24}$$

where $c > 0$ is an absolute constant. The right hand side of (5.24) is small if the unknown vector $\boldsymbol{\mu}$ can be well approximated by a piecewise linear sequence in $\mathcal{S}_n^{\cup}$ with not too many pieces.

The leading constant in (5.24) is 6. We will show that sparsity pattern aggregation achieves a substantially better performance. We obtain the sharp oracle inequality (5.27) below, improving upon (5.24) not only in the fact that the leading constant is 1 but also in the rate of the remainder term; we will see that the exponent 5/4 of the logarithmic factor is reduced to 1.

For any set $J \subseteq \{2, ..., n-1\}$, define

$$\nu_J := \frac{\exp(-|J|)}{H_C \binom{n-2}{|J|}}, \qquad H_C := \sum_{i=0}^{n-2} \exp(-i). \tag{5.25}$$

Let $Q_J \in \mathbb{R}^{n \times n}$ be the projector on the linear subspace W_J of $\mathbb{R}^n$ given by

$$W_J := \Big\{ \boldsymbol{u} \in \mathbb{R}^n : \ \forall i \in \{2, ..., n-1\} \setminus J, \ 2u_i = u_{i+1} + u_{i-1} \Big\}.$$

Given a vector $\mathbf{y}$ of observations and $\boldsymbol{\theta} = (\theta_J)_{J \subseteq \{2,...,n-1\}}$ where each θ_J belongs to $\mathbb{R}$, let

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \sum_{J \subseteq \{2,...,n-1\}} \theta_J Q_J \mathbf{y}.$$

Finally, let

$$\hat{\boldsymbol{\mu}}^{Q-conv} = \boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}}$$

where $\hat{\boldsymbol{\theta}}$ is the solution of the optimization problem

$$\min_{\boldsymbol{\theta} \in \Lambda'} \ \|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \mathbf{y}\|^2 + \sum_{J \subset \{2,...,n-1\}} \theta_J \left(\frac{2\sigma^2 |J|}{n} + \frac{1}{2} \|\boldsymbol{\mu}_{\boldsymbol{\theta}} - Q_J \mathbf{y}\|^2 + \frac{46\sigma^2}{n} \log \frac{1}{\nu_J} \right)$$

where

$$\Lambda' = \left\{ \boldsymbol{\theta} : \ \theta_J \geq 0 \text{ for all } J \subseteq \{2, ..., n-1\}, \text{ and } \sum_{J \subseteq \{2,...,n-1\}} \theta_J = 1 \right\}.$$

The structure of this minimization problem is the same as of its analog introduced in Section 5.2. This is a quadratic program that aggregates the linear estimators $(Q_J \mathbf{y})_{J \subseteq \{2,...,n-1\}}$ using the Q-aggregation procedure [32, 33, 11] with the prior weights (5.25).

Theorem 5.6. *Let $\boldsymbol{\mu} \in \mathbb{R}^n$, $n \geq 3$, and assume that the noise vector $\boldsymbol{\xi}$ has distribution $\mathcal{N}(0, \sigma^2 I_{n \times n})$. There exist absolute constants $c, c' > 0$ such that for all $\delta \in (0, 1/3)$, the estimator $\hat{\boldsymbol{\mu}}^{Q-conv}$ satisfies with probability at least $1 - 3\delta$,*

$$\|\hat{\boldsymbol{\mu}}^{Q-conv} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathbb{R}^n} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c\sigma^2 q(\boldsymbol{u})}{n} \log \frac{en}{q(\boldsymbol{u})} \right) + \frac{c\sigma^2 \log(1/\delta)}{n}, \tag{5.26}$$

and we have

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{Q-conv} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathbb{R}^n} \left(\|\boldsymbol{\mu} - \boldsymbol{u}\|^2 + \frac{c'\sigma^2 q(\boldsymbol{u})}{n} \log \frac{en}{q(\boldsymbol{u})} \right). \tag{5.27}$$

The proof of this theorem is the same as that of Theorem 5.1 with the only difference that J is now a subset of $\{2, ..., n-1\}$ rather than that of $\{1, ..., n-1\}$, and we replace the notation P_J and V_J by Q_J and W_J respectively.

The leading constant of the oracle inequality (5.27) is 1, and the remainder term is proportional to $q(\boldsymbol{u}) \log(en/q(\boldsymbol{u}))$. These are two improvements upon (5.24), where the leading constant is 6 and the remainder term is proportional to $q(\boldsymbol{u}) \log(en/q(\boldsymbol{u}))^{5/4}$.

In view of (5.27), for the class of piecewise linear convex sequences with at most q linear pieces, $\mathcal{S}_n^{\cup}(q) = \{\boldsymbol{u} \in \mathcal{S}_n^{\cup} : q(\boldsymbol{u}) \le q\}$ we have the following bounds for the maximal expected regrets

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\| - \min_{\boldsymbol{u} \in \mathcal{S}_n^{\cup}(q)} \|\boldsymbol{u} - \boldsymbol{\mu}\| \right) \le c \sqrt{\frac{\sigma^2 q}{n} \log\left(\frac{en}{q}\right)}, \tag{5.28}$$

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E} \|\hat{\boldsymbol{\mu}}^Q - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u} \in \mathcal{S}_n^{\cup}(q)} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \right) \le \frac{c\sigma^2 q}{n} \log\left(\frac{en}{q}\right), \tag{5.29}$$

where $c > 0$ is an absolute constant. The same bounds hold for the minimax risks over $\mathcal{S}_n^{\cup}(q)$ since the minimax risk is smaller than the minimax regret.

The following proposition shows that the rates of convergence in (5.28) and (5.29) are optimal up to logarithmic factors. We omit the discussion since it is similar to that after Proposition 5.4.

Proposition 5.7. *Let $n \ge 3$. There exist absolute constants $c, c' > 0$ such that, for any positive integer $q \le n$,*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}(q)} \mathbb{P}_{\boldsymbol{\mu}} \left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \ge \frac{c\sigma^2 q}{n} \right) > c',$$

where the infimum is taken over all estimators.

Proof. Assume that $q \ge 2$ since for $q = 1$ the result is trivial. We also assume for simplicity that n is a multiple of q. Let $m = n/q$ and $\gamma = (1/8)\sqrt{\sigma^2 q/n}$. Set $\beta_0 = 0, \alpha_0 = 0$ and define, for all integers $j \ge 1$,

$$\beta_j = \beta_{j-1} + \gamma + m\alpha_{j-1}, \qquad \alpha_j = 2\gamma + \alpha_{j-1}. \tag{5.30}$$

By the Varshamov-Gilbert bound [99, Lemma 2.9] there exists $\Omega \subset \{0,1\}^q$ such that (5.22) is satisfied, with k replaced by q. For each $\boldsymbol{\omega} \in \Omega$, define a vector $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathbb{R}^n$ with components

$$u^{\boldsymbol{\omega}}_{jm+i} = \omega_{j+1}\gamma + \alpha_j(i-1) + \beta_j, \qquad j = 0, ..., q-1, \quad i = 1, ..., m.$$

The sequence $\boldsymbol{u}^{\boldsymbol{\omega}}$ is piecewise linear. It is linear with slope α_j on the set $\{jm+1, ..., (j+1)m\}$ for any $j = 0, ..., q-1$. Thus, $q(\boldsymbol{u}^{\boldsymbol{\omega}}) = q$. Next, we prove that $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathcal{S}_n^{\cup}$ for all $\boldsymbol{\omega} \in \Omega$. It is enough to check the convexity condition at the endpoints of the linear pieces:

$$2u^{\boldsymbol{\omega}}_{jm} \le u^{\boldsymbol{\omega}}_{jm-1} + u^{\boldsymbol{\omega}}_{jm+1}, \qquad 2u^{\boldsymbol{\omega}}_{jm+1} \le u^{\boldsymbol{\omega}}_{jm} + u^{\boldsymbol{\omega}}_{jm+2}, \tag{5.31}$$

for all $j = 1, ..., q-1$. Using (5.30) we get that, for all $j = 1, ..., q-1$,

$$\begin{aligned} u^{\boldsymbol{\omega}}_{jm+1} - u^{\boldsymbol{\omega}}_{jm} &= \omega_{j+1}\gamma + \beta_j - (\omega_j\gamma + \alpha_{j-1}(m-1) + \beta_{j-1}), \\ &= (\omega_{j+1} - \omega_j + 1)\gamma + \alpha_{j-1}, \\ &= (\omega_{j+1} - \omega_j - 1)\gamma + \alpha_j. \end{aligned}$$

Hence, $\alpha_{j-1} \le u^{\boldsymbol{\omega}}_{jm+1} - u^{\boldsymbol{\omega}}_{jm} \le \alpha_j$. Since also $\alpha_{j-1} = u^{\boldsymbol{\omega}}_{jm} - u^{\boldsymbol{\omega}}_{jm-1}$ and $\alpha_j = u^{\boldsymbol{\omega}}_{jm+2} - u^{\boldsymbol{\omega}}_{jm+1}$, it follows that the two inequalities (5.31) hold, for all $j = 1, ..., q-1$. Thus, $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathcal{S}_n^{\cup}$. In summary, we have proved that $\boldsymbol{u}^{\boldsymbol{\omega}} \in \mathcal{S}_n^{\cup}(q)$ for all $\boldsymbol{\omega} \in \Omega$.

Now, from the Varshamov-Gilbert bound, cf. (5.22), for $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$ we have

$$\|\boldsymbol{u}^{\boldsymbol{\omega}} - \boldsymbol{u}^{\boldsymbol{\omega}'}\|^2 = \frac{\gamma^2}{q} d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') \ge \frac{\gamma^2}{8} = \frac{\sigma^2 q}{512n},$$

where $d_H(\cdot, \cdot)$ is the Hamming distance. Finally, similarly to (5.23), the Kullback-Leibler divergence between $P_{\boldsymbol{\omega}}$ and $P_{\mathbf{0}}$ satisfies $K(P_{\boldsymbol{\omega}}, P_{\mathbf{0}}) \le \frac{\log(|\Omega|-1)}{16}$. Applying [99, Theorem 2.7] with $\alpha = 1/16$ completes the proof. □

5.4 Concluding remarks and discussion

In this short note, we have shown that the estimators $\hat{\boldsymbol{\mu}}^{Q}$ and $\hat{\boldsymbol{\mu}}^{Q-conv}$ based on sparsity pattern aggregation (in its Q-aggregation version) achieve oracle inequalities that improve on some previous results for isotonic and convex regression.

One of the improvements is that oracle inequalities (5.10) and (5.27) are sharp, i.e., with leading constant 1 and they are valid for all $\boldsymbol{\mu} \in \mathbb{R}^n$. It allows us to obtain bounds for the minimax regret under arbitrary model misspecification, which was not possible based on the previous results. We show that these bounds are rate optimal up to logarithmic factors. The question on whether the least squares estimators under monotonicity and convexity constraints can achieve sharp oracle inequalities with correct rates remains open.

Another improvement is that we obtain oracle inequalities both with high probability and in expectation, which was not the case in the previous work.

An advantage of the least squares estimator is that it requires no tuning parameters. In particular, the knowledge of σ^2 is not needed to construct the estimators $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\uparrow})$ and $\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\cup})$. This is in contrast to the ℓ_1 penalized estimator (5.4) and the estimators $\hat{\boldsymbol{\mu}}^{Q}$ and $\hat{\boldsymbol{\mu}}^{Q-conv}$; their construction requires the knowledge of σ^2. For the ℓ_1 penalized estimator (5.4), the issue may be addressed by using a scale-free version of the Lasso [16, 95]. For the Q-aggregation estimators $\hat{\boldsymbol{\mu}}^{Q}$ and $\hat{\boldsymbol{\mu}}^{Q-conv}$, we can treat the issue of unknown σ as in [11]. Namely, it is shown in [11] that the oracle inequalities for Q-aggregation procedures are essentially preserved after plugging in an estimator $\hat{\sigma}^2$ of σ^2 that satisfies $|\hat{\sigma}^2/\sigma^2 - 1| \le 1/8$ with high probability, which is even weaker than consistency.

Finally, note that instead of Q-aggregation we could have used sparsity pattern aggregation by the Exponential Screening procedure of [91]. This would lead to sharp oracle inequalities in expectation of the form (5.10) and (5.27) but not to inequalities with high probability such as (5.9) and (5.26). This is the reason why we have opted for Q-aggregation rather than for Exponential Screening in this paper. On the other hand, Exponential Screening estimators are computationally more attractive than Q-aggregation since they can be successfully approximated by MCMC algorithms (see [91, 93] for details).

Part III

Shape restricted regression

Chapter 6

Sharp oracle inequalities for Least Squares estimators in shape restricted regression

The performance of Least Squares estimators is studied in shape restricted regression for convex cones that include nondecreasing sequences, convex sequences and higher order cones. We derive sharp oracle inequalities for the Least Squares estimator, i.e., oracle inequalities with leading constant 1. Two types of oracle inequalities are derived. The inequalities of the first type are adaptive in the sense that the rate becomes parametric if the true sequence can be well approximated by a sequence that satisfies some low-dimensionality property. The inequalities of the second type yield a rate that corresponds to the nonparametric rate of smoothness classes under a localized Gaussian width assumption. The oracle inequalities hold in deviation with exponential probability bounds and in expectation. To obtain our results, we improve the best known bounds on the statistical dimension of the cone of convex sequences, and we derive upper bounds on the statistical dimension of higher order cones. Then we construct an estimator that aggregates two projections on opposite convex cones. In isotonic regression, the estimator adapts to the best direction of monotonicity. In convex regression, the estimator mimics the best behavior among concavity and convexity. Our estimators are fully data-driven and computationally tractable.

6.1 Introduction

Assume that we have the observations

$$Y_i = \mu_i + \xi_i, \qquad i = 1, ..., n, \tag{6.1}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathbb{R}^n$ is unknown, $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$ is a noise vector with n-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ where $\sigma > 0$ and $I_{n\times n}$ is the $n \times n$ identity matrix. Denote by $\mathbb{E}_{\boldsymbol{\mu}}$ the expectation with respect to the distribution of the random variable $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\xi}$. The vector $\mathbf{y} = (Y_1, ..., Y_n)^T$ is observed and the goal is to estimate $\boldsymbol{\mu}$. The estimation error is measured with the scaled norm $\|\cdot\|$ defined by

$$\|\boldsymbol{u}\|^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2, \qquad \boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n.$$

The error of an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ is given by $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$. Let also $|\cdot|_2^2$ be the squared Euclidean norm, so that $\frac{1}{n}|\cdot|_2^2 = \|\cdot\|^2$.

Let E be a subset of $\mathbb{R}^n$. If the unknown regression vector $\boldsymbol{\mu}$ lies in E, we say that the model is well-specified. If $\boldsymbol{\mu} \in E$, an estimator $\hat{\boldsymbol{\mu}}$ enjoys good performance if the squared error

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \tag{6.2}$$

is small, either in expectation or with high probability. If $\boldsymbol{\mu} \notin E$, we say that the model is misspecified. In that case, the quantity of interest to assess the performance of an estimator $\hat{\boldsymbol{\mu}}$ with respect to the set E is the regret

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u} \in E} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2. \tag{6.3}$$

The Least Squares estimator over a closed convex set $\mathcal{K}$ is defined by

$$\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}) \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathcal{K}} \|\mathbf{y} - \boldsymbol{u}\|^2.$$

If V is a linear subspace of $\mathbb{R}^n$ of dimension d_V, then $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(V)$ is the orthogonal projection of $\mathbf{y}$ onto V and the Pythagorean theorem yields

$$\forall \boldsymbol{\mu} \in \mathbb{R}^n, \qquad \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(V) - \boldsymbol{\mu}\|^2 = \min_{\boldsymbol{u} \in V} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 d_V}{n}. \tag{6.4}$$

In this paper, we are interested in the performance of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ when $\mathcal{K}$ is closed and convex and we propose generalizations of (6.4).

If $\mathcal{K}$ is a closed convex set, there are general methods to bound (6.2) from above if $\boldsymbol{\mu} \in \mathcal{K}$ and $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$. For example, the analysis of [29] ensures that (6.2) is bounded from above by $t_{\sigma,n}(\mathcal{K})^2/n$ with high probability if

$$\mathbb{E}_{\boldsymbol{\mu}} \sup_{\boldsymbol{u} \in \mathcal{K} : |\boldsymbol{u} - \boldsymbol{\mu}|_2 \le t_{\sigma,n}(\mathcal{K})} \boldsymbol{\xi}^T (\boldsymbol{u} - \boldsymbol{\mu}) \le \frac{t_{\sigma,n}(\mathcal{K})^2}{2}, \tag{6.5}$$

where $t_{\sigma,n}(\mathcal{K})$ is a constant that may depend on σ, n and $\mathcal{K}$. The left hand side of (6.5) is sometimes called the localized Gaussian width of $\mathcal{K}$. However, if $\boldsymbol{\mu} \notin \mathcal{K}$, to our knowledge there is no general method to control the regret (6.3) with $E = \mathcal{K}$ using conditions similar to (6.5) on the localized Gaussian width. One of the goals of the present paper is to fill this gap. Theorem 6.12 allows us to derive inequalities of the form

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}) - \boldsymbol{\mu}\|^2 \le \min_{\boldsymbol{u} \in \mathcal{K}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{c t_{\sigma,n}(\mathcal{K})^2}{n}, \tag{6.6}$$

in expectation and with high probability for some absolute constant $c > 0$, as soon as

$$\mathbb{E}_{\boldsymbol{\mu}} \sup_{\boldsymbol{u} \in \mathcal{K} : |\boldsymbol{u} - \Pi_{\mathcal{K}}(\boldsymbol{\mu})|_2 \le t_{\sigma,n}(\mathcal{K})} \boldsymbol{\xi}^T (\boldsymbol{u} - \Pi_{\mathcal{K}}(\boldsymbol{\mu})) \le \frac{t_{\sigma,n}(\mathcal{K})^2}{2}$$

is satisfied, where $\Pi_{\mathcal{K}}(\boldsymbol{\mu})$ is the projection of $\boldsymbol{\mu}$ onto $\mathcal{K}$. The inequality (6.6) is a sharp oracle inequality, i.e., an oracle inequality with leading constant 1, which yields an upper bound on the regret (6.3) with E replaced by $\mathcal{K}$. This is opposed to other oracle inequalities of the form

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \le C \min_{\boldsymbol{u} \in \mathcal{K}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{t_{\sigma,n}(\mathcal{K})^2}{n}, \tag{6.7}$$

where the leading constant C is strictly greater than 1. The right hand sides of (6.6) and (6.7) are minimized if $\boldsymbol{u}$ is the projection of $\boldsymbol{\mu}$ onto $\mathcal{K}$, i.e., $\min_{\boldsymbol{u}\in\mathcal{K}} \|\boldsymbol{u}-\boldsymbol{\mu}\|^2 = \|\Pi_{\mathcal{K}}(\boldsymbol{\mu})-\boldsymbol{\mu}\|^2$ where $\Pi_{\mathcal{K}} : \mathbb{R}^n \to \mathcal{K}$ is the projection onto $\mathcal{K}$.

If $\mathcal{K}$ is a closed convex cone, we will show that the estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ satisfies oracle inequalities of the form

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2 \le C \min_{\boldsymbol{u}\in\mathcal{K}} \left(\|\boldsymbol{u}-\boldsymbol{\mu}\|^2 + r_{\sigma,n}(\boldsymbol{u})\right), \tag{6.8}$$

where $C \ge 1$ and the remainder term $r_{\sigma,n}(\boldsymbol{u})$ depends on $\boldsymbol{u}$, unlike (6.7) where the remainder term depends on $\mathcal{K}$ but not on $\boldsymbol{u}$. The right hand side of (6.8) is minimized by a vector $\boldsymbol{u} \in \mathcal{K}$ that makes a trade-off between the approximation error $\|\boldsymbol{u}-\boldsymbol{\mu}\|^2$ and the quantity $r_{\sigma,n}(\boldsymbol{u})$. Bounds such as (6.8) will be called adaptive oracle inequalities. Such trade-off is common in the context of sparse linear regression, where the right hand side of sparsity oracle inequalities balances approximation error and sparsity [17, 61, 92]. This is opposed to the oracle inequality (6.7) where there is no trade-off, the right hand side of (6.7) is always minimized for $\boldsymbol{u} = \Pi_{\mathcal{K}}(\boldsymbol{\mu})$ which achieves the smallest approximation error.

6.1.1 Preliminary properties of closed convex sets

We recall here several properties of convex sets that will be used in the paper. Given a closed convex set $\mathcal{K} \subset \mathbb{R}^n$, denote by $\Pi_{\mathcal{K}} : \mathbb{R}^n \to \mathcal{K}$ the projection onto $\mathcal{K}$. For all $\mathbf{y} \in \mathbb{R}^n$, $\Pi_{\mathcal{K}}(\mathbf{y})$ is the unique vector in $\mathcal{K}$ such that

$$(\boldsymbol{u}-\Pi_{\mathcal{K}}(\mathbf{y}))^T(\mathbf{y}-\Pi_{\mathcal{K}}(\mathbf{y})) \le 0, \qquad \boldsymbol{u}\in\mathcal{K}. \tag{6.9}$$

Inequality (6.9) can be rewritten as follows

$$\|\Pi_{\mathcal{K}}(\mathbf{y})-\mathbf{y}\|^2 + \|\boldsymbol{u}-\Pi_{\mathcal{K}}(\mathbf{y})\|^2 \le \|\boldsymbol{u}-\mathbf{y}\|^2, \qquad \mathbf{y}\in\mathbb{R}^n, \boldsymbol{u}\in\mathcal{K}, \tag{6.10}$$

which is a consequence of the cosine theorem. The Least Squares estimator over $\mathcal{K}$ is exactly the projection of $\mathbf{y}$ onto $\mathcal{K}$, i.e., $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}) = \Pi_{\mathcal{K}}(\mathbf{y})$. In this case, (6.10) yields that for all $\boldsymbol{u} \in \mathcal{K}$,

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})-\mathbf{y}\|^2 \le \|\boldsymbol{u}-\mathbf{y}\|^2 - \|\boldsymbol{u}-\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})\|^2. \tag{6.11}$$

Inequality (6.11) can be interpreted in terms of strong convexity: the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ solves an optimization problem where the function to minimize is strongly convex with respect to the norm $\|\cdot\|$. Strong convexity grants inequality (6.11), which is stronger than the inequality

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})-\mathbf{y}\|^2 \le \|\boldsymbol{u}-\mathbf{y}\|^2, \qquad \boldsymbol{u}\in\mathcal{K}.$$

Now, assume that $\mathcal{K}$ is a closed convex cone. For all $\mathbf{y} \in \mathbb{R}^n$, $\Pi_{\mathcal{K}}(\mathbf{y})$ is the unique vector in $\mathcal{K}$ that satisfies

$$\Pi_{\mathcal{K}}(\mathbf{y})^T\mathbf{y} = |\Pi_{\mathcal{K}}(\mathbf{y})|_2^2 \quad \text{and} \quad \forall \boldsymbol{\theta}\in\mathcal{K}, \ \boldsymbol{\theta}^T\mathbf{y} \le \boldsymbol{\theta}^T\Pi_{\mathcal{K}}(\mathbf{y}). \tag{6.12}$$

The lineality space of the closed convex cone $\mathcal{K}$ is the linear space

$$\mathrm{Lin}(\mathcal{K}) = \{\boldsymbol{u}\in\mathcal{K} : \quad -\boldsymbol{u}\in\mathcal{K}\}.$$

It is the maximal linear subspace contained in $\mathcal{K}$. Not all cones admit a lineality space different than $\{\mathbf{0}\}$, for instance $\mathrm{Lin}(\mathbb{R}^{n+}) = \{\mathbf{0}\}$ where $\mathbb{R}^{n+}$ is the nonnegative orthant. For any $\boldsymbol{v} \in \mathbb{R}^n$ we have

$$|\Pi_{\mathcal{K}}(\boldsymbol{v})|_2^2 = \boldsymbol{v}^T \Pi_{\mathcal{K}}(\boldsymbol{v}) = \left(\sup_{\boldsymbol{\theta} \in \mathcal{K} : |\boldsymbol{\theta}|_2 \leq 1} \boldsymbol{v}^T \boldsymbol{\theta} \right)^2, \tag{6.13}$$

cf. [1, Proposition 3.1 and Appendix B.4]. Define

$$\delta(\mathcal{K}) := \mathbb{E}\left[|\Pi_{\mathcal{K}}(\boldsymbol{g})|_2^2 \right] = \mathbb{E}\left[\boldsymbol{g}^T \Pi_{\mathcal{K}}(\boldsymbol{g}) \right] = \mathbb{E}\left[\left(\sup_{\boldsymbol{\theta} \in \mathcal{K} : |\boldsymbol{\theta}|_2 \leq 1} \boldsymbol{g}^T \boldsymbol{\theta} \right)^2 \right], \tag{6.14}$$

where $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, I_{n \times n})$. We will refer to the quantity $\delta(\mathcal{K})$ as the statistical dimension of the cone $\mathcal{K}$. We refer the reader to [1] for properties and equivalent definitions of $\delta(\cdot)$. The statistical dimension appears in several results of the present paper. The following proposition is a first generalization of (6.4).

Proposition 6.1. *Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\mathcal{K}$ be a closed convex subset of $\mathbb{R}^n$. Define the cone $\mathcal{T}_{K, \Pi_K(\boldsymbol{\mu})} = \{t(\boldsymbol{v} - \Pi_K(\boldsymbol{\mu})) : \, t \geq 0, \boldsymbol{v} \in K\}$. Then*

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in K} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2}{n} \delta\left(\mathcal{T}_{K, \Pi_K(\boldsymbol{\mu})} \right). \tag{6.15}$$

Proof. Let $\boldsymbol{u} = \Pi_K(\boldsymbol{\mu})$ and $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)$. Then (6.11) yields

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \leq (2/n) \boldsymbol{\xi}^T (\hat{\boldsymbol{\mu}} - \boldsymbol{u}) - \|\hat{\boldsymbol{\mu}} - \boldsymbol{u}\|^2. \tag{6.16}$$

Using the simple inequality $2ab - b^2 \leq a^2$ and taking the supremum, we obtain

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \leq \frac{1}{n} \left(\sup_{\boldsymbol{v} \in K, \boldsymbol{v} \neq \boldsymbol{u}} \frac{\boldsymbol{\xi}^T (\boldsymbol{v} - \boldsymbol{u})}{|\boldsymbol{v} - \boldsymbol{u}|_2} \right)^2.$$

Combining the expectation of the previous display and (6.14) completes the proof. □

In the well-specified case, a similar upper bound was derived in [84, Theorem 3.1]. Oymak and Hassibi [84] also proved a worst-case lower bound that matches the upper bound. If $K \subset V$ where V is a subspace of dimension d_V, then by monotonicity of the statistical dimension (cf. [1, Proposition 3.1]) $\delta(\mathcal{T}_{K, \Pi_K(\boldsymbol{\mu})}) \leq \delta(V) = d_V$. In this case, (6.15) shows that the constant 4 in [96, Proposition 3.1] can be reduced to 1.

6.1.2 Contributions and organisation of the paper

Section 6.2 presents several examples of closed convex cones and a review of the literature on the performance of the Least Squares estimator over these cones. The contributions of the present paper are the following.

First, we present in Proposition 6.3 a link between the performance of the Least Squares estimator over a closed convex cone and the statistical dimensions of certain cones. Section 6.3 allows us to derive several oracle inequalities of the form (6.8), where $C = 1$ and $\hat{\boldsymbol{\mu}}$ is the Least Squares estimator over a closed convex cone. These oracle inequalities are given in Theorems 6.2, 6.6, 6.8 and 6.10.

Second, we develop a new argument to bound from above the statistical dimension of the cones defined in Example 6.2 and Example 6.3 below. The result is given

in Theorem 6.5 and Theorem 6.7. For the cones of convex sequences, this is a substantial improvement upon (6.29).

Third, Section 6.4 provides a general technique to obtain sharp oracle inequalities similar to (6.6). Theorem 6.12 describes this result. In Theorem 6.12, we also explain why this result is of a different nature than the recent analysis of [29].

Finally, Section 6.5 deals with the problem of adaptation to the direction of monotonicity. Consider the mean estimation problem (6.1) under monotonicity constraint, either nondecreasing or non-increasing. We could not find in the literature a data-driven procedure that automatically mimics the best estimator among the two Least Squares estimators $\{\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}), \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\downarrow})\}$, where $\mathcal{S}_n^{\downarrow} = -\mathcal{S}_n^{\uparrow}$ is the cone of non-increasing sequences. The same question arises in the context of univariate convex regression: it is possible to mimic the best estimator among $\{\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}), \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{S}_n^{\cup})\}$? The procedure presented in Section 6.5 gives a positive answer. In particular, we show that it is possible to aggregate two Least Squares estimators on two opposite cones.

Section 6.7 is devoted to the proofs.

6.2 Examples of closed convex cones

We give below several examples of sets K that will be studied in the paper. For all $q \geq 2$, denote by D_q the following matrix with $q-1$ rows and q columns

$$D_q := \begin{bmatrix} -1 & 1 & 0 & \dots & \dots & 0 \\ 0 & -1 & 1 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & -1 & 1 \end{bmatrix}. \tag{6.17}$$

For all $q \geq 2$, denote by $\leq$ and $\geq$ the component-wise comparison operators in $\mathbb{R}^q$, i.e.,

$$(u_1, ..., u_q)^T \leq (v_1, ..., v_q)^T \quad \text{if and only if} \quad \forall i = 1, ..., q, \quad u_i \leq v_i.$$

Example 6.1 (Nondecreasing sequences)**.** Let $\mathcal{S}_n^{\uparrow}$ be the set of all nondecreasing sequences, defined by

$$\begin{aligned} \mathcal{S}_n^{\uparrow} &:= \{\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n : u_i \leq u_{i+1}, \quad i = 1, ..., n-1\}, \\ &:= \{\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n : D_n \boldsymbol{u} \geq \mathbf{0} = (0, ..., 0)^T\}, \end{aligned}$$

where D_n is the matrix (6.17). Define the matrix $\mathbb{X} = (\mathbb{X}_{ij})_{i=1,...,n,\ j=1,...,n}$ by

$$\mathbb{X}_{ij} = 1 \quad \text{if } j \leq i \qquad \text{and} \qquad \mathbb{X}_{ij} = 0 \quad \text{otherwise.} \tag{6.18}$$

Then

$$\mathcal{S}_n^{\uparrow} = \left\{\mathbb{X}\boldsymbol{\theta}, \qquad \boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T \in \mathbb{R}^n : \quad \theta_k \geq 0 \text{ for all } k \geq 2\right\}. \tag{6.19}$$

The set $\mathcal{S}_n^{\uparrow}$ is a closed convex cone. An exact formula is available for the statistical dimension of $\mathcal{S}_n^{\uparrow}$. Namely, it is proved in [1, (D.12)] that

$$\delta(\mathcal{S}_n^{\uparrow}) = \sum_{k=1}^{n} \frac{1}{k}. \tag{6.20}$$

For $\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathcal{S}_n^{\uparrow}$, let $k(\boldsymbol{u}) \geq 1$ be the integer such that $k(\boldsymbol{u}) - 1$ is the number of inequalities $u_i \leq u_{i+1}$ that are strict for $i = 1, \dots, n-1$ (the number of jumps of $\boldsymbol{u}$). If $\boldsymbol{u} = \mathbb{X}\boldsymbol{\theta}$ with $\boldsymbol{\theta}$ as in (6.19), then $k(\boldsymbol{u})$ is also the number of strictly positive entries among $\theta_2, ..., \theta_n$. The cone $\mathcal{S}_n^{\uparrow}$ is endowed with the the lineality space

$$\mathrm{Lin}(\mathcal{S}_n^{\uparrow}) = \{(b, ..., b)^T, \quad b \in \mathbb{R}\} = \{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow} : k(\boldsymbol{u}) = 1\},$$

which is the subspace of constant sequences. Previous results on the performance of the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ can be found in [80, 107, 27, 29], where risk bounds or oracle inequalities with leading constant strictly greater than 1 are derived. Two types of risk bounds or oracle inequalities have been obtained so far. If, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathcal{S}_n^{\uparrow}$, it is known [80, 107, 27, 29] that for some absolute constant $c > 0$,

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \frac{c\sigma^2 \log(en)}{n} + c\left(\frac{(\mu_n - \mu_1)\sigma^2}{n}\right)^{2/3}. \tag{6.21}$$

If $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$, the following oracle inequality was proved in [27]:

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq 6 \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right). \tag{6.22}$$

The assumption $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ is rather restrictive as it does not allow for any model misspecification. We will see below that this assumption can be dropped. If $D > 0$ is a fixed parameter and $\log(en)^3\sigma^2 \leq nD^2$, the bound (6.21) yields the rate $(D\sigma^2)^{2/3}n^{-2/3}$ for the risk of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$. By the lower bound [14, Corollary 5], this rate is minimax optimal over the class $\{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow} : \mu_n - \mu_1 \leq D\}$ if $\log(en)^3\sigma^2 \leq nD^2$. The bound (6.22) yields the rate $n^{-2/3}$ up to logarithmic factors, thanks to the approximation argument given in [14, Lemma 2]. The oracle inequality (6.22) also yields a parametric rate (up to logarithmic factors) if $\boldsymbol{\mu}$ is well approximated by a piecewise constant sequence with not too many pieces.

Example 6.2 (Convex sequences)**.** If $n \geq 3$, define the set of convex sequences $\mathcal{S}_n^{\cup}$ by

$$\begin{aligned}\mathcal{S}_n^{\cup} &\coloneqq \{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : \ 2u_i \leq u_{i+1} + u_{i-1}, \ i = 2, \dots, n-1\},\\ &\coloneqq \{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : \ D_{n-1}D_n\boldsymbol{u} \geq \boldsymbol{0} = (0, ..., 0)^T\},\end{aligned}$$

where D_{n-1} and D_n are the rectangular matrices defined in (6.17). If $\mathbb{X}$ is the matrix defined in (6.18), then

$$\mathcal{S}_n^{\cup} = \left\{\mathbb{X}^2\boldsymbol{\theta}, \qquad \boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T \in \mathbb{R}^n : \quad \theta_k \geq 0 \text{ for all } k \geq 3\right\}. \tag{6.23}$$

If $x_1 < ... < x_n$ are equispaced design points in $\mathbb{R}$, i.e., $x_i = (i-1)(x_2 - x_1) + x_1$, $i = 2, ..., n$, then

$$\mathcal{S}_n^{\cup} = \{\boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{u} = (f(x_1), ..., f(x_n))^T \text{ for some convex function } f : \mathbb{R} \to \mathbb{R}\}.$$

If $x_1 < ... < x_n$ are non-equispaced design points in $\mathbb{R}$, define the cone

$$\mathcal{K}^C_{x_1,\dots,x_n} \coloneqq \{\boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{u} = (f(x_1), ..., f(x_n))^T \text{ for some convex function } f : \mathbb{R} \to \mathbb{R}\},$$

which can be rewritten as

$$\mathcal{K}^C_{x_1,...,x_n} := \{\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n : \quad \frac{u_i - u_{i-1}}{x_i - x_{i-1}} \leq \frac{u_{i+1} - u_i}{x_{i+1} - x_i}, i = 2, ..., n-1\}. \tag{6.24}$$

For any $\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathcal{K}^C_{x_1,...,x_n}$, we say that $\boldsymbol{u}$ is piecewise affine with k pieces if there exist real numbers $a_1, ..., a_k$ and a partition $(T_1, ..., T_k)$ of $\{1, .., n\}$ such that

$$u_i = a_j(x_i - x_l) + u_l, \qquad i, l \in T_j, \qquad j = 1, ..., k. \tag{6.25}$$

If $\boldsymbol{u} = (f(x_1), ..., f(x_n))^T$ for some convex function $f : \mathbb{R} \to \mathbb{R}$ and f is a piecewise affine function with k pieces, then $\boldsymbol{u}$ is piecewise affine with k pieces. For any $\boldsymbol{u} \in \mathcal{K}^C_{x_1,...,x_n}$, let $q(\boldsymbol{u}) \geq 1$ be the smallest integer such that $\boldsymbol{u}$ is piecewise affine with $q(\boldsymbol{u})$ pieces. The quantity $q(\boldsymbol{u}) \geq 1$ satisfies

$$q(\boldsymbol{u}) - 1 \leq \left| \left\{ i = 2, ..., n-1 : \quad \frac{u_i - u_{i-1}}{x_i - x_{i-1}} < \frac{u_{i+1} - u_i}{x_{i+1} - x_i} \right\} \right|.$$

If $\boldsymbol{u} \in \mathcal{S}_n^{\cup}$ and $\boldsymbol{u} = \mathbb{X}^2\boldsymbol{\theta}$ with $\boldsymbol{\theta} \geq \mathbf{0}$ as in (6.23), then $q(\boldsymbol{u}) - 1 \leq |\{i = 3, ..., n : \theta_i > 0\}|$. The cone $\mathcal{K}^C_{x_1,...,x_n}$ is endowed with the lineality space

$$\mathrm{Lin}(\mathcal{K}^C_{x_1,...,x_n}) = \{(ax_1 + b, ..., ax_n + b)^T, a, b \in \mathbb{R}\} = \{\boldsymbol{u} \in \mathcal{K}^C_{x_1,...,x_n} : q(\boldsymbol{u}) = 1\},$$

which is the subspace of all affine sequences. The performance of the Least Squares estimator over convex sequences has been recently studied in [50, 27], where it was proved that if $\boldsymbol{\mu} \in \mathcal{K}^C_{x_1,...,x_n}$, the estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq C \left(\min_{\boldsymbol{u} \in \mathcal{S}_n^{\cup}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 q(\boldsymbol{u})}{n} \left(\log \frac{en}{q(\boldsymbol{u})} \right)^{5/4} \right) \right). \tag{6.26}$$

If $\boldsymbol{\mu} \in \mathcal{K}^C_{x_1,...,x_n}$ and $nR^2_{\boldsymbol{\mu}} \geq \log(en)^{5/4}\sigma^2$ where $R_{\boldsymbol{\mu}}$ is defined in Corollary 6.14 below, then the estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq C \left(\frac{\sqrt{R_{\boldsymbol{\mu}}}\sigma^2}{n} \right)^{4/5} \log(en), \tag{6.27}$$

where $C > 0$ is a constant that depends only on the ratio

$$\frac{\max_{i=2,...,n}(x_i - x_{i-1})}{\min_{i=2,...,n}(x_i - x_{i-1})}. \tag{6.28}$$

The bound (6.26) yields an almost parametric rate if $\boldsymbol{\mu}$ can be well approximated by a piecewise affine sequence with not too many pieces. If $\bar{R} > 0$ is a fixed parameter and $n\bar{R}^2 \geq \log(en)^{5/4}\sigma^2$, the bound (6.27) yields the rate $(\bar{R}^2\sigma^8)^{1/5}n^{-4/5}$, which is minimax optimal up to logarithmic factors [50]. We will see that the assumption $\boldsymbol{\mu} \in \mathcal{K}^C_{x_1,...,x_n}$ can be dropped and that the bounds above can be improved by deriving the corresponding sharp oracle inequalities (cf. (6.43) and (6.61)). It is not known whether (6.27) or (6.26) holds for some absolute constant $C > 0$ independent of the design points. We will prove in Theorem 6.6 below a more general version of (6.26)

that holds for any design points $x_1 < ... < x_n$. The following upper bound on the statistical dimension of the cone $\mathcal{K}^C_{x_1,...,x_n}$ is derived in [50]:

$$\delta(\mathcal{K}^C_{x_1,...,x_n}) \leq c(\log(en))^{5/4}, \tag{6.29}$$

for some constant $c > 0$ that depends on the ratio (6.28). A tighter bound will be derived in Theorem 6.5.

Example 6.3 (β-constrained sequences)**.** For any positive integer $\beta < n$, consider the weights vector $\boldsymbol{\omega}^{[\beta]} = (\omega_0^{[\beta]}, ..., \omega_\beta^{[\beta]})^T \in \mathbf{R}^{\beta+1}$ defined by

$$\omega_k^{[\beta]} = (-1)^{\beta-k}\binom{\beta}{k}, \qquad k = 0, ..., \beta. \tag{6.30}$$

Define the cone

$$\begin{aligned}\mathcal{S}_n^{[\beta]} &:= \{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : \ (\boldsymbol{\omega}^{[\beta]})^T(u_i, u_{i+1}, ..., u_{i+\beta})^T \geq 0, \ i = 1, \dots, n-\beta\}, \\ &:= \{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : \ D_{n-\beta+1}D_{n-\beta+2}...D_n\boldsymbol{u} \geq \mathbf{0} = (0, ..., 0)^T\},\end{aligned}$$

where $D_{n-\beta+1}, ..., D_n$ are the matrices defined in (6.17). In particular, we have $\boldsymbol{\omega}^{[1]} = (-1, 1)^T$ and $\mathcal{S}^{[1]} = \mathcal{S}_n^{\uparrow}$ is the cone of nondecreasing sequences, $\boldsymbol{\omega}^{[2]} = (1, -2, 1)^T$ and $\mathcal{S}^{[2]} = \mathcal{S}_n^{\cup}$ is the cone of convex sequences. The notation β has been chosen to highlight the similarity between the cones $\mathcal{S}_n^{[\beta]}$ and β-smoothness classes in nonparametric statistics. In univariate regression for instance, the minimax rate of estimation under the loss (6.2) for smoothness classes such as Hölder or Sobolev balls is proportional to $n^{-2\beta/(2\beta+1)}$, where β is the smoothness of the class. Inequalities (6.21) and (6.27) show that the rates of convergence of the Least Squares estimator over the cones $\mathcal{S}_n^{[\beta]}$ under the loss (6.2) are $n^{-2\beta/(2\beta+1)}$ up to logarithmic factors for $\beta = 1, 2$.

If $\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathcal{S}_n^{[\beta]}$, we say that $\boldsymbol{u}$ is a piecewise polynomial function of degree d with k pieces if there exist polynomials $Q_1, ..., Q_k$ of degree at most d and a partition $(T_1, ..., T_k)$ of $\{1, ..., n\}$ such that

$$u_i = Q_j(i), \qquad i \in T_j, \qquad j = 1, ..., k. \tag{6.31}$$

If $\boldsymbol{u} \in \mathcal{S}_n^{[\beta]}$, define $s_\beta(\boldsymbol{u}) \geq 1$ as the smallest integer s such that $\boldsymbol{u}$ is a piecewise polynomial function of degree $\beta - 1$ with s pieces. Note that $s_1(\cdot) = k(\cdot)$ for nondecreasing sequences, and $s_2(\cdot) = q(\cdot)$ for convex sequences. The cone $\mathcal{S}_n^{[\beta]}$ is endowed with with the lineality space

$$\mathrm{Lin}(\mathcal{S}_n^{[\beta]}) = \{\boldsymbol{u} \in \mathcal{S}_n^{[\beta]} : s_\beta(\boldsymbol{u}) = 1\},$$

which is the subspace of polynomials of degree at most $\beta - 1$. In Section 6.7.1 we will derive upper bounds on the statistical dimension of the cones $\mathcal{S}_n^{[\beta]}$ for all $\beta \geq 2$. These bounds lead to sharp oracle inequalities for $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{[\beta]})$ similar to (6.22) and (6.26).

Example 6.4 (m-monotone sequences)**.** Define the linear operator $\nabla : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\nabla\boldsymbol{u} = (u_2 - u_1, ..., u_n - u_{n-1}, -u_n), \qquad \boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n.$$

If $m \geq 0$ is an integer, the cone of m-monotone sequences is defined as

$$\mathcal{M}_n^m = \{\boldsymbol{u} \in \mathbb{R}^n : \nabla^m\boldsymbol{u} \geq 0\}.$$

For density estimation, m-monotone functions have been studied in [5, 6]. Simple algebra shows that $\mathcal{M}_n^0 = \mathbb{R}^{n+}$ and $\mathcal{M}_n^1 = \mathcal{S}_n^{\uparrow} \cap (-\mathbb{R}^{n+})$, where $\mathbb{R}^{n+}$ is the nonnegative orthant. For $m = 2$, $\mathcal{M}_n^2 = \mathcal{S}_n^{\cup} \cap (-\mathcal{S}_n^{\uparrow}) \cap \mathbb{R}^{n+}$ is the cone of convex, non-increasing and nonnegative sequences. For all $m \geq 1$, we have

$$\mathcal{M}_n^m = \cap_{l=0}^m \left((-1)^{m-l} \mathcal{S}_n^{[l]} \right), \tag{6.32}$$

with the convention $\mathcal{S}_n^{[0]} = \mathbb{R}^{n+}$. This implies that $\mathcal{M}_n^m \subset (-1)^{m-1} \mathcal{S}_n^{\uparrow}$. Using the monotonicity of the statistical dimension with respect to inclusion [1, Proposition 3.1], we obtain $\delta(\mathcal{M}_n^m) \leq \delta((-1)^{m-1}\mathcal{S}_n^{\uparrow}) = \delta(\mathcal{S}_n^{\uparrow}) \leq \log(en)$. This simple monotonicity argument cannot be used to bound from above the statistical dimension of the cones $\mathcal{S}_n^{\cup}$ or $\mathcal{S}_n^{[\beta]}$ defined above. For all $m \geq 0$, the cone $\mathcal{M}_n^m$ contains no linear subspace and

$$\mathrm{Lin}(\mathcal{M}_n^m) = \{\mathbf{0}\}.$$

This is opposed to the cones $\mathcal{S}_n^{[\beta]}$, since the lineality space of $\mathcal{S}_n^{[\beta]}$ has dimension β.

Example 6.5 (Cone with arbitrary weights vector $\boldsymbol{\omega}$)**.** Given a positive integer m with $m \leq n$ and a vector of weights $\boldsymbol{\omega} \in \mathbb{R}^m$, define the cone $\mathcal{K}_n^{\boldsymbol{\omega}}$ by

$$\mathcal{K}_n^{\boldsymbol{\omega}} = \Big\{ \boldsymbol{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n : \ \boldsymbol{\omega}^T (u_i, u_{i+1}, ..., u_{i+m-1})^T \geq 0 \\ \text{for all } i = 1, \ldots, n-m+1 \Big\}.$$

The cone $\mathcal{K}_n^{\boldsymbol{\omega}}$ is a closed and convex subset of $\mathbb{R}^n$. The sets $\mathcal{S}_n^{\uparrow}$, $\mathcal{S}_n^{\cup}$ and $\mathcal{S}_n^{[\beta]}$ defined above are examples of cones of this form.

Example 6.6 (Polyhedral cones)**.** All the examples above are particular cases of convex polyhedral cones. Given a matrix A with r rows and n columns, define the cone

$$\mathcal{C}_A = \{\boldsymbol{u} \in \mathbb{R}^n, \quad A\boldsymbol{u} \leq \mathbf{0} = (0, ..., 0)^T\},$$

where $\leq$ denotes the component wise comparison in $\mathbf{R}^r$. The polyhedral cones in $\mathbb{R}^n$ are the sets of the form $\mathcal{C}_A$ where A is a matrix with n columns. The lineality space of $\mathcal{C}_A$ is the kernel of matrix A. The cone of nondecreasing sequences and the cone of convex sequences are polyhedral cones with $\mathcal{S}_n^{\uparrow} = \mathcal{C}_{-D_n}$ and $\mathcal{S}_n^{\cup} = \mathcal{C}_{-D_{n-1}D_n}$. For cones of higher order defined in Example 6.3, we have $\mathcal{S}_n^{[\beta]} = \mathcal{C}_{A_\beta}$ where

$$A_\beta = -D_{n-\beta+1} D_{n-\beta+2} ... D_n.$$

Given a vector $\boldsymbol{\omega} \in \mathbb{R}^m$ as in Example 6.5, the cone $\mathcal{K}_n^{\boldsymbol{\omega}}$ satisfies $\mathcal{K}_n^{\boldsymbol{\omega}} = \mathcal{C}_{A_{\boldsymbol{\omega}}}$ where $A_{\boldsymbol{\omega}} = (a_{ij})_{i=1,...,n; j=1,...,n-m+1}$ is the matrix

$$a_{ij} = -\omega_{j+i-1} \quad \text{if } 1 \leq i+j-1 \leq m \qquad \text{and} \qquad a_{ij} = 0 \quad \text{otherwise.}$$

6.3 Sharp oracle inequalities and adaptation

6.3.1 Nondecreasing sequences

Let us start this section with the following result on the performance of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$. This result is representative of the sharp oracle inequalities obtained for different cones in the rest of the section.

Theorem 6.2. *For all $n \geq 2$ and any $\boldsymbol{\mu} \in \mathbb{R}^n$,*

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right). \tag{6.33}$$

Furthermore, for any $a > 0$ and any $t > 0$, we have

$$\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{(1+a)\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right) + \frac{\sigma^2(8 + 2/a)t}{n} \tag{6.34}$$

with probability greater than $1 - \exp(-t)$.

The proof will be given in the next section. Let us discuss some features of Theorem 6.2 that are new. First, the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ satisfies oracle inequalities both in deviation with exponential probability bounds and in expectation, cf. (6.34) and (6.33), respectively. Previously known oracle inequalities for the Least Squares estimator over $\mathcal{S}_n^{\uparrow}$ were only proved in expectation.

Second, both (6.33) and (6.34) are sharp oracle inequalities, i.e., with leading constant 1. Although sharp oracle inequalities were obtained using aggregation methods [14], this is the first known sharp oracle inequality for the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$.

Third, the assumption $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ is not needed, as opposed to the result of [27].

Last, the constant 1 in front of $\frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})}$ in (6.33) is optimal for the Least Squares estimator. To see this, assume that there exists an absolute constant $c < 1$ such that for all $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ and $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$,

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{c\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right). \tag{6.35}$$

Set $\boldsymbol{\mu} = 0$. Thanks to (6.20), the left hand side of the above display becomes $\sigma^2 \sum_{k=1}^n 1/k$ which is greater than $\sigma^2 \log(n)$, while the right hand side becomes $c\sigma^2 \log(en)$. Thus, it is impossible to improve the constant in front of $\frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})}$ for the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$. However, it is still possible that for another estimator $\hat{\boldsymbol{\mu}}$, (6.35) holds with $c < 1$ or without the logarithmic factor. We do not know whether such an estimator exists.

We now highlight the adaptive behavior of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$. Let $\boldsymbol{u}^* \in \mathcal{S}_n^{\uparrow}$ be a minimizer of the right hand side of (6.33). Let $k = k(\boldsymbol{u}^*)$ and let $T_1, ..., T_k$ be a partition of $\{1, ..., n\}$ such that $\boldsymbol{u}^*$ is constant on all T_j, $j = 1, .., k$. Given $T_1, ..., T_k$, consider the piecewise constant oracle

$$\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}} \in \operatorname*{argmin}_{\boldsymbol{u} \in W_{T_1,...,T_k}} \|\mathbf{y} - \boldsymbol{u}\|^2,$$

where $W_{T_1,...,T_k}$ is the linear subspace of all sequences that are constant on all T_j, $j = 1, ..., k$. This subspace has dimension k, so the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}}$ satisfies

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}} - \boldsymbol{\mu}\|^2 = \min_{\boldsymbol{u} \in W_{T_1,...,T_k}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k}{n} \leq \|\boldsymbol{u}^* - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k}{n}.$$

Thus, (6.33) can be interpreted in the sense that without the knowledge of $T_1, ..., T_k$, the performance of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ is similar to that of $\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}}$ up to the factor $\log(en/k)$. Of course, the knowledge of $T_1, ..., T_k$ is not accessible in practice, so $\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}}$ is an oracle that can only serve as a benchmark. This adaptive behavior of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ was observed in [27].

6.3.2 Orthogonal decomposition and lineality spaces

Proposition 6.3 below is our main tool to derive sharp oracle inequalities for the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ for any closed convex cone $\mathcal{K}$. Given a matrix P, denote by $\operatorname{Im} P$ the linear span of the columns of P.

Proposition 6.3. *Let $n \geq 2$, $\boldsymbol{\mu} \in \mathbb{R}^n$, let $\mathcal{K}$ be a closed convex set and let $\boldsymbol{u} \in \mathcal{K}$. Furthermore, assume (i) and (ii) below.*

(i) There exist orthogonal projectors $P_1, ..., P_k$ such that

$$\sum_{j=1}^{k} P_j = I_{n\times n} \qquad \text{and} \qquad P_j P_l = 0 \quad \text{for all } j, l = 1, ..., k.$$

(ii) There exist closed convex cones $\mathcal{K}_1, ..., \mathcal{K}_k$ such that

$$\{P_j \boldsymbol{v}, \boldsymbol{v} \in \mathcal{K}\} \subseteq \mathcal{K}_j \subseteq \operatorname{Im} P_j \quad \text{and} \quad P_j \boldsymbol{u} \in \operatorname{Lin}(\mathcal{K}_j), \qquad j = 1, ..., k.$$

Then almost surely

$$\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{K}) - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \sum_{j=1}^{k} \|\Pi_j(P_j \boldsymbol{\xi})\|^2, \tag{6.36}$$

where $\Pi_j : \operatorname{Im} P_j \to \mathcal{K}_j$ is the projection onto $\mathcal{K}_j$, $j = 1, ..., k$.

The assumptions of Proposition 6.3 on the projectors $P_1, ..., P_k$ imply

$$\operatorname{Im} P_1 \oplus \cdots \oplus \operatorname{Im} P_k = \mathbb{R}^n,$$

where $\oplus$ denotes an orthogonal direct sum. The random variables $P_1\boldsymbol{\xi}, ..., P_k\boldsymbol{\xi}$ are thus independent normal random variables.

Proposition 6.3 allows us to bound from above the loss of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ by bounding from above the sum of the k independent random variables $\|\Pi_1(P_1\boldsymbol{\xi})\|^2, ..., \|\Pi_k(P_k\boldsymbol{\xi})\|^2$. By the definition of the statistical dimension of a cone given in (6.14), Proposition 6.3 shows that upper bounds on the statistical dimensions of the cones $\mathcal{K}_1, ..., \mathcal{K}_k$ imply a sharp oracle inequality for the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$.

If $\mathcal{K}$ is a closed convex cone, a natural choice for $\mathcal{K}_1, ..., \mathcal{K}_k$ is $\mathcal{K}_j = \{P_j \boldsymbol{v}, \boldsymbol{v} \in \mathcal{K}\}$, $j = 1, ..., k$. However, we will see in Theorem 6.10 an application of Proposition 6.3 with a different choice for the cones $\mathcal{K}_1, ..., \mathcal{K}_k$, and in (6.41) an application of Proposition 6.3 where $\mathcal{K}$ is not a cone.

To prove Proposition 6.3, we use the strong convexity of the Least Squares minimization problem, as follows.

Proof of Proposition 6.3. Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{LS}(\mathcal{K})$ for notational simplicity. Inequality (6.16) can be rewritten as

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 - |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 \leq 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u}) - |\boldsymbol{u} - \hat{\boldsymbol{\mu}}|_2^2 = |\boldsymbol{\xi}|_2^2 - |\boldsymbol{\xi} - \hat{\boldsymbol{\mu}} + \boldsymbol{u}|_2^2. \tag{6.37}$$

For all $j = 1, ..., k$, we have $P_j\hat{\boldsymbol{\mu}} \in \mathcal{K}_j$ and $P_j\boldsymbol{u} \in \mathrm{Lin}(\mathcal{K}_j)$, so $P_j(\hat{\boldsymbol{\mu}} - \boldsymbol{u}) \in \mathcal{K}_j$. Thus, by definition of $\Pi_1, ..., \Pi_k$,

$$\begin{aligned}
|\boldsymbol{\xi}|_2^2 - |\boldsymbol{\xi} - \hat{\boldsymbol{\mu}} + \boldsymbol{u}|_2^2 &= \sum_{j=1}^{k} |P_j\boldsymbol{\xi}|_2^2 - |P_j(\boldsymbol{\xi} - (\hat{\boldsymbol{\mu}} - \boldsymbol{u}))|_2^2, \\
&\leq \sum_{j=1}^{k} |P_j\boldsymbol{\xi}|_2^2 - |P_j\boldsymbol{\xi} - \Pi_j(P_j\boldsymbol{\xi})|_2^2, \\
&= \sum_{j=1}^{k} 2(P_j\boldsymbol{\xi})^T\Pi_j(P_j\boldsymbol{\xi}) - |\Pi_j(P_j\boldsymbol{\xi})|_2^2 = \sum_{j=1}^{k} |\Pi_j(P_j\boldsymbol{\xi})|_2^2,
\end{aligned}$$

where for the last equality we used that if Π is a projection onto a closed convex cone, $\Pi(\boldsymbol{\theta})^T(\boldsymbol{\theta} - \Pi(\boldsymbol{\theta})) = 0$ for all vectors $\boldsymbol{\theta}$ (cf. (6.12)). By plugging the previous display back into (6.37) and dividing by n, we obtain (6.36). □

For any $T \subset \{1, ..., n\}$ and $\boldsymbol{v} \in \mathbb{R}^n$, denote by $\boldsymbol{v}_T \in \mathbb{R}^{|T|}$ the restriction of $\boldsymbol{v}$ to the set T and by $|T|$ the cardinality of T. Let $(T_1, ..., T_k)$ be a partition of $\{1, ..., n\}$ and let $P_1, ..., P_k$ be the coordinate projections

$$P_j = \sum_{l \in T_j} \boldsymbol{e}_l\boldsymbol{e}_l^T, \qquad \text{for all} \quad j = 1, ..., k. \tag{6.38}$$

If $\mathcal{K} = \mathcal{K}_n^{\boldsymbol{\omega}}$ for some vector $\boldsymbol{\omega} \in \mathbb{R}^m$ (cf. Example 6.5), and the cones $\mathcal{K}_1, ..., \mathcal{K}_k$ are given by $\mathcal{K}_j = P_j\mathcal{K}$ for all $j = 1, ..., k$, then $\mathcal{K}_j = \mathcal{K}_{|T_j|}^{\boldsymbol{\omega}}$ and Proposition 6.3 takes the following form.

Corollary 6.4. *Let $\boldsymbol{\omega} \in \mathbb{R}^m$ for some $m \leq n$. Let $(T_1, ..., T_k)$ be a partition of $\{1, ..., n\}$ such that for all $j = 1, ..., k$, T_j has the form $T_j = \{t_j + 1, ..., t_j + |T_j|\}$ for some integer $t_j \geq 0$. Let $\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathcal{K}_n^{\boldsymbol{\omega}}$ be such that*

$$\boldsymbol{u}_{T_j} \in \mathrm{Lin}(\mathcal{K}_{|T_j|}^{\boldsymbol{\omega}}), \qquad j = 1, ..., k.$$

Then, almost surely

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}_n^{\boldsymbol{\omega}}) - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \sum_{j=1}^{k} \|\Pi_{\mathcal{K}_{|T_j|}^{\boldsymbol{\omega}}}(\boldsymbol{\xi}_{T_j})\|^2. \tag{6.39}$$

To illustrate Proposition 6.3 and Corollary 6.4, we now prove Theorem 6.2.

Proof of Theorem 6.2. Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}_n^{\uparrow})$ for notational simplicity. Let $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$ and let $k = k(\boldsymbol{u})$. Let $T_1, ..., T_k$ be a partition of $\{1, ..., n\}$ such that $\boldsymbol{u}$ is constant on all T_j, $j = 1, ..., k$. Thanks to Corollary 6.4 with $\boldsymbol{\omega} = (-1, 1)^T$, inequality (6.39) holds where $\mathcal{K}_{|T_j|}^{\boldsymbol{\omega}} = \mathcal{S}_{|T_j|}^{\uparrow}$. We will first prove (6.33). It was shown in [1, Appendix D.4] that if $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, I_{n\times n})$, then

$$\mathbb{E}\left[|\Pi_{\mathcal{S}_{|T|}^{\uparrow}}(\boldsymbol{g}_T)|_2^2\right] = \sum_{l=1}^{|T|} \frac{1}{l} \leq \log(e|T|), \tag{6.40}$$

where $T = \{1, ..., n\}$. To complete the proof of (6.33), we take the expectations in (6.39) and apply Jensen's inequality to get

$$\sum_{j=1}^{k} \log(e|T_j|) = k\sum_{j=1}^{k} \frac{1}{k}\log(e|T_j|) \leq k\log\left(\frac{e}{k}\sum_{j=1}^{k}|T_j|\right) = k\log\frac{en}{k}.$$

To prove (6.34), we use (6.39) where $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$ is a minimizer of the right hand side of (6.34), and then apply Lemma 6.20 to the stochastic term. □

As a consequence of Proposition 6.3, we can derive sharp oracle inequalities for the Least Squares estimator if we can bound from above the statistical dimension of certain cones. The survey [1] provides general recipes to bound from above the statistical dimension of cones of several types. For instance, the statistical dimension of $\mathcal{S}_n^\uparrow$ is given by the exact formula (6.20). Bounds on the statistical dimension of a closed convex cone $\mathcal{K}$ can be based on metric entropy results, as $\sigma^2\delta(\mathcal{K})/n = \mathbb{E}_{\mathbf{0}}\|\Pi_{\mathcal{K}}(\boldsymbol{\xi})\|^2$ is the risk of the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ when the true vector is $\mathbf{0}$. This technique is used in [50] to derive the bound (6.29).

We now illustrate how Proposition 6.3 can be used in situations where $\mathcal{K}$ is not a cone. Let $\mathcal{K}$ be any closed convex subset of $\mathcal{S}_n^\uparrow$. Let $\boldsymbol{u} \in \mathcal{K}$ and let $(T_1, ..., T_k)$ be a partition of $\{1, ..., n\}$ such that $\boldsymbol{u}$ is constant on all $T_j, j = 1, ..., k$. Let $P_1, ..., P_k$ be the coordinate projections (6.38) and let $\mathcal{K}_j = \mathcal{S}_{|T_j|}^\uparrow$. Applying (6.36) and following the same arguments as in the proof of Theorem 6.2, we obtain that for any closed convex subset $\mathcal{K}$ of $\mathcal{S}_n^\uparrow$,

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{K}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{K}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right). \tag{6.41}$$

For instance, (6.41) holds for $\mathcal{K} = \{\boldsymbol{u} \in \mathcal{S}_n^\uparrow : a^- \leq u_1, u_n \leq a^+\}$ where $a^- < a^+$ are fixed real numbers.

6.3.3 Convex sequences and arbitrary design

We now present a new argument to bound from above the statistical dimension of the cone of convex sequences.

Theorem 6.5. *Let $n \geq 3$. Let $x_1 < ... < x_n$ be real numbers and consider the cone $\mathcal{K}^C_{x_1,...,x_n}$ defined in (6.24). Let $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$. Then*

$$\delta(\mathcal{K}^C_{x_1,...,x_n}) = \mathbb{E}|\Pi_{\mathcal{K}^C_{x_1,...,x_n}}(\boldsymbol{g})|_2^2 \leq 10 \log(en). \tag{6.42}$$

The proof of Theorem 6.5 is given in Section 6.7.1. The bound (6.42) improves upon (6.29) as the exponent 5/4 is reduced to 1. Furthermore, (6.42) does not depend on the design points $x_1, ..., x_n$. Combining Proposition 6.3 and Theorem 6.5 yields the following oracle inequalities.

Theorem 6.6. *Let $n \geq 3$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. Let $x_1 < ... < x_n$ be real numbers and consider the cone $\mathcal{K}^C_{x_1,...,x_n}$ defined in (6.24). Then*

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{K}^C_{x_1,...,x_n}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{K}^C_{x_1,...,x_n}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{10\sigma^2 q(\boldsymbol{u})}{n} \log \frac{en}{q(\boldsymbol{u})} \right). \tag{6.43}$$

Furthermore, for any $t > 0$ we have

$$\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{K}^C_{x_1,...,x_n}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{K}^C_{x_1,...,x_n}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{20\sigma^2 q(\boldsymbol{u})}{n} \log \frac{en}{q(\boldsymbol{u})} \right) + \frac{10\sigma^2 t}{n} \tag{6.44}$$

with probability greater than $1 - \exp(-t)$.

Proof of Theorem 6.6. Let $\boldsymbol{u} \in \mathcal{K}^C_{x_1,...,x_n}$ and let $k = q(\boldsymbol{u})$. Let $T_1, ..., T_k$ be a partition of $\{1, ..., n\}$ such that $\boldsymbol{u}$ is affine on all T_j, $j = 1, ..., k$ (cf. (6.25)). Thanks to

Proposition 6.3 with the projectors $P_1, ..., P_k$ defined in (6.38), inequality (6.36) holds with $\mathcal{K}_j = \mathcal{K}^C_{x_{i_1},...,x_{i_{|T_j|}}}$ and $T_j = \{i_1, ..., i_{|T_j|}\}$. The rest of the proof is identical to that of Theorem 6.2 with $a = 1$, except that we use the bound (6.42) instead of (6.40). □

The oracle inequalities of Theorem 6.6 do not depend on the design points $x_1, ..., x_n$. In particular, (6.43) and (6.44) hold for non-equispaced design points and design points that are be arbitrarily close to each other. This improves upon the oracle inequality (6.26) proved in [50, 27] where C is strictly greater than 1 and depends on the design points through the ratio (6.28). The sharp oracle inequalities of Theorem 6.6 hold in deviation with exponential probability bounds and in expectation for any $\boldsymbol{\mu} \in \mathbb{R}^n$, whereas previously known oracle inequalities from [50, 27] only hold in expectation under the additional assumption that $\boldsymbol{\mu} \in \mathcal{K}^C_{x_1,...,x_n}$.

6.3.4 Minimax regret bounds for $\mathcal{S}^{[\beta]}_n$

The argument behind Theorem 6.5 can be used to recursively control the statistical dimensions of the cones $\mathcal{S}^{[\beta]}_n$ for $\beta \geq 3$.

Theorem 6.7. *Let β, n be integers such that $1 \leq \beta < n$. Then*

$$\delta(\mathcal{S}^{[\beta]}_n) = \mathbb{E}|\Pi_{\mathcal{S}^{[\beta]}_n}(\boldsymbol{g})|_2^2 \leq C(\beta)\log(en), \tag{6.45}$$

where $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$ and $C(\beta) = 3 \cdot 4^{\beta-1} - 2$.

The proof of Theorem 6.7 is given in Section 6.7.1. We now generalize Theorem 6.2 to the cones $\mathcal{S}^{[\beta]}_n$ for $\beta \geq 1$.

Theorem 6.8. *Let β, n be integers such that $1 \leq \beta < n$ and let $\boldsymbol{\mu} \in \mathbb{R}^n$. Then*

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{[\beta]}_n) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{S}^{[\beta]}_n}\left(\|\boldsymbol{u}-\boldsymbol{\mu}\|^2 + \frac{C(\beta)\sigma^2 s_\beta(\boldsymbol{u})}{n}\log\frac{en}{s_\beta(\boldsymbol{u})}\right), \tag{6.46}$$

where $C(\beta)$ depends only on β. Furthermore, for any $t > 0$ we have

$$\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{S}^{[\beta]}_n) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{S}^{[\beta]}_n}\left(\|\boldsymbol{u}-\boldsymbol{\mu}\|^2 + \frac{2C(\beta)\sigma^2 s_\beta(\boldsymbol{u})}{n}\log\frac{en}{s_\beta(\boldsymbol{u})}\right) + \frac{10\sigma^2 t}{n}$$

with probability greater than $1 - \exp(-t)$.

Proof of Theorem 6.8. Let $\boldsymbol{u} \in \mathcal{S}^{[\beta]}_n$ and let $k = s_\beta(\boldsymbol{u})$. Let $T_1, ..., T_k$ be a partition of $\{1, ..., n\}$ such that $\boldsymbol{u}$ is a polynomial of degree $\beta - 1$ on all T_j, $j = 1, ..., k$. We apply Corollary 6.4 with $\boldsymbol{\omega} = \boldsymbol{\omega}^{[\beta]}$, where $\boldsymbol{\omega}^{[\beta]}$ is defined in (6.30). Then inequality (6.39) holds with $\mathcal{K}^{\boldsymbol{\omega}}_{|T_j|} = \mathcal{S}^{[\beta]}_{|T_j|}$. The rest of the proof is the same as the proof of Theorem 6.2 with $a = 1$, except that we use the bound (6.45) instead of (6.40). □

For $\beta = 1$, the result above is exactly Theorem 6.2 with $a = 1$. The following lower bound holds.

Theorem 6.9. *There exists an absolute constant $c > 0$ such that the following holds. Let β, s, n be positive integers such that $n \geq s$. Then*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{[\beta]}:\, s_\beta(\boldsymbol{\mu}) \leq s} \mathbb{P}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq \frac{c(\beta) s}{n} \right) \geq c,$$

where the infimum is taken over all estimators and $c(\beta) > 0$ is a constant that depends only on β.

The proof of Theorem 6.9 is given in Section 6.7.2. Under the assumption of Theorem 6.9 for the integers s, β and n, Markov inequality yields

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{[\beta]}:\, s_\beta(\boldsymbol{\mu}) \leq s} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq \frac{c(\beta) c s}{n}. \tag{6.47}$$

Consider the class

$$\mathcal{S}_n^{[\beta]}(s) := \{\boldsymbol{\mu} \in \mathcal{S}_n^{[\beta]} : s_\beta(\boldsymbol{\mu}) \leq s\}. \tag{6.48}$$

The left hand side of (6.47) is the minimax risk over this class. We have proved that the minimax risk over this class is of the order $\sigma^2 s/n$, up to a logarithmic factor. To be more precise, inequalities (6.46) and (6.47) yield

$$\frac{c(\beta) c \sigma^2 s}{n} \leq \inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{[\beta]}(s)} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \frac{C(\beta) \sigma^2 s \log(en/s)}{n}.$$

Define the minimax regret as

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u} \in \mathcal{S}_n^{[\beta]}(s)} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \right).$$

Since the oracle inequality (6.46) is sharp, thus it implies the following bound on the maximal expected regret of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{[\beta]})$

$$\sup_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{[\beta]}) - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u} \in \mathcal{S}_n^{[\beta]}(s)} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \right) \leq \frac{C(\beta) \sigma^2 s \log(en/s)}{n}. \tag{6.49}$$

Since the minimax risk is always smaller than the minimax regret, the minimax regret also satisfies

$$\frac{c(\beta) c \sigma^2 s}{n} \leq \inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \mathbb{R}^n} \left(\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u} \in \mathcal{S}_n^{[\beta]}(s)} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \right) \leq \frac{C(\beta) \sigma^2 s \log(en/s)}{n}. \tag{6.50}$$

For $\beta = 1, 2$, this bracketing of the minimax regret was shown in [14]. The estimator proposed in [14] for which the upper bound of (6.50) is attained is an aggregate of an exponential number of estimators, and cannot be computed in polynomial time. In (6.50), the upper bound on the minimax regret is attained at the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{[\beta]})$, which can be computed efficiently by solving a convex quadratic minimization program of size n.

6.3.5 Cones of m-monotone sequences

This section deals with the cones defined in Example 6.4. Unlike the cones $\mathcal{S}_n^{\uparrow}, \mathcal{S}_n^{\cup}$ and $\mathcal{S}_n^{[\beta]}$ studied in the previous sections, the lineality space of the cone $\mathcal{M}_n^m$ is $\{\mathbf{0}\}$ for all $n, m \geq 1$.

Let $(T_1, ..., T_k)$ be a partition of $\{1, ..., n\}$. To derive Theorem 6.2, we applied Corollary 6.4 to the cones $\mathcal{S}_{|T_1|}^{\uparrow}, ..., \mathcal{S}_{|T_k|}^{\uparrow}$ and each of these cones has a lineality space of dimension 1. Similarly, to derive Theorem 6.8 we applied Corollary 6.4 to the cones $\mathcal{S}_{|T_1|}^{[\beta]}, ..., \mathcal{S}_{|T_k|}^{[\beta]}$ and each of these cones has a lineality space of dimension β. Although the lineality space of the cone $\mathcal{M}_n^m$ is $\{\mathbf{0}\}$ for all $m, n \geq 1$, Proposition 6.3 can be used to derive the following oracle inequalities.

Theorem 6.10. *Let m, n be integers such that $1 \leq m < n$ and let $\boldsymbol{\mu} \in \mathbb{R}^n$. Then*

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{M}_n^m) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{M}_n^m} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{C(m)\sigma^2 s_m(\boldsymbol{u})}{n} \log \frac{en}{s_m(\boldsymbol{u})} \right),$$

where $C(\cdot)$ is the constant from Theorem 6.7. Furthermore, for any $t > 0$,

$$\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{M}_n^m) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{M}_n^m} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{2C(m)\sigma^2 s_m(\boldsymbol{u})}{n} \log \frac{en}{s_m(\boldsymbol{u})} \right) + \frac{10\sigma^2 t}{n}$$

with probability greater than $1 - \exp(-t)$.

For $\boldsymbol{u} \in \mathcal{M}_n^m$, recall that $s_m(\boldsymbol{u})$ is the smallest number s such that $\boldsymbol{u}$ is a piecewise polynomial function of degree at most $m - 1$ with s pieces (cf. (6.31)).

Proof of Theorem 6.10. Let $\boldsymbol{u} \in \mathcal{M}_n^m$ and let $k = s_m(\boldsymbol{u})$. Let $T_1, ..., T_k$ be a partition of $\{1, ..., n\}$ such that $\boldsymbol{u}$ is a polynomial of degree $m - 1$ on all T_j, $j = 1, ..., k$. Let $P_1, ..., P_k$ be the coordinate projections (6.38). We apply Proposition 6.3 to the cones $\mathcal{K}_j := \mathcal{S}_{|T_j|}^{[m]}$, $j = 1, ..., k$. For all $j = 1, ..., k$, $P_j\mathcal{K} \subset \mathcal{K}_j$ (cf. (6.32)) and the restriction $\boldsymbol{u}_{T_j}$ is a polynomial of degree at most m. Thus, $\boldsymbol{u}_{T_j} \in \mathrm{Lin}(\mathcal{K}_j)$. Then inequality (6.36) holds and can be rewritten in the form

$$\|\hat{\boldsymbol{\mu}}^{LS}(\mathcal{M}_n^m) - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \sum_{j=1}^{k} \|\Pi_{\mathcal{S}_{|T_j|}^{[m]}}(\boldsymbol{\xi}_{T_j})\|^2.$$

The rest of the proof is the same as the proof of Theorem 6.2 with $a = 1$, except that instead of (6.40) we use the bound (6.45) with β replaced by m. □

6.3.6 Non-Gaussian noise

In this section, we do not assume that the noise vector $\boldsymbol{\xi}$ is normally distributed. Proposition 6.3 does not depend on the distribution of the noise vector $\boldsymbol{\xi}$. To illustrate this, we apply Corollary 6.4 to the cone $\mathcal{K} = \mathcal{S}_n^{\uparrow}$. For any $\boldsymbol{u} \in \mathbb{R}^n$, let $(T_1, ..., T_k)$ be a partition such that $\boldsymbol{u}$ is constant on all $T_j, j = 1, ..., k$. Taking expectations of both sides of (6.39) yields

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \sum_{j=1}^{k} \mathbb{E}\|\Pi_{\mathcal{S}_{|T_j|}^{\uparrow}}(\boldsymbol{\xi}_{T_j})\|^2. \tag{6.51}$$

Let $\boldsymbol{\xi} = (\varepsilon_1, ..., \varepsilon_N)^T$ where $\varepsilon_1, ..., \varepsilon_N$ are i.i.d. random variables with $\mathbb{E}\varepsilon_1 = 0$ and $\mathbb{E}[\varepsilon_1^2] \leq \sigma^2$. It was shown in [27, Theorem 3.1] that for all $N \geq 1$,

$$\mathbb{E}\|\Pi_{\mathcal{S}_N^\uparrow}(\boldsymbol{\xi})\|^2 \leq 4\sigma^2 \log(eN).$$

Combining this bound with (6.51) and Jensen's inequality yields the following result.

Corollary 6.11. *Let $\boldsymbol{\mu} \in \mathbb{R}^n$. Let $\boldsymbol{\xi} = (\varepsilon_1, ..., \varepsilon_n)^T$ where $\varepsilon_1, ..., \varepsilon_N$ are i.i.d. with $\mathbb{E}\varepsilon_1 = 0$ and $\mathbb{E}[\varepsilon_1^2] = \sigma^2$. Then for all $\boldsymbol{u} \in \mathcal{S}_n^\uparrow$,*

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^\uparrow) - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{4\sigma^2 k(\boldsymbol{u}) \log(en/k(\boldsymbol{u}))}{n}.$$

6.3.7 Multivariate isotonic regression

Proposition 6.3 is not limited to univariate regression. Let $d > 1$, let $n_1, ..., n_d > 1$ be integers and let $n = n_1 n_2 ... n_d$. Consider the discrete hyperrectangle

$$I = \{(i_1, ..., i_d) \subset \mathbf{N}^d, \quad 1 \leq i_l \leq n_l, \text{ for all } l = 1, ..., d\}, \tag{6.52}$$

and the cone $\mathcal{K}^{d\uparrow} \subset \mathbf{R}^I$ defined by

$$\mathcal{K}^{d\uparrow} = \Big\{\boldsymbol{u} = (u_{i_1 i_2 ... i_d})_{(i_1, i_2, ..., i_d) \in I} \in \mathbf{R}^I, \\ \text{such that } u_{i_1 i_2 ... i_d} \leq u_{j_1 j_2 ... j_d} \text{ if } (i_l \leq j_l \text{ for all } l = 1, ..., d)\Big\}. \tag{6.53}$$

The set $\mathcal{K}^{d\uparrow}$ is the cone of vectors indexed by I that are nondecreasing in all directions $l = 1, ..., d$. For $d = 2$, the performance of the Least Squares estimator over $\mathcal{K}^{2\uparrow}$ has been recently studied in [28]. Let k be a positive integer. Consider a partition $(T_1, ..., T_k)$ of I and $\boldsymbol{u} \in \mathcal{K}^{d\uparrow}$ such that $\boldsymbol{u}$ is constant on T_j for all $j = 1, ..., k$. Then, for any unknown $\boldsymbol{\mu} \in \mathbf{R}^I$, Proposition 6.3 yields

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{K}^{d\uparrow}) - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2}{n}\sum_{j=1}^{k} \delta_j, \tag{6.54}$$

where for all $j = 1, ..., r$, δ_j is the statistical dimension of the cone

$$\Big\{\boldsymbol{u} = (u_{i_1 i_2 ... i_d})_{(i_1, i_2, ..., i_d) \in T_j} \in \mathbf{R}^{T_j} \\ \text{such that } u_{i_1 i_2 ... i_d} \leq u_{j_1 j_2 ... j_d} \text{ if } (i_l \leq j_l \text{ for all } l = 1, ..., d)\Big\}.$$

If $d = 2$ and $T_1, ..., T_k$ are rectangles, Chatterjee et al. [28] proved that $\delta_j \leq C \log(en)^8$ for some absolute constant C. In that case, a direct consequence of Proposition 6.3 is

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{K}^{2\uparrow}) - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{C\sigma^2 k \log(en)^8}{n}.$$

This improves upon the oracle inequality [28, Theorem 4.1] that has a leading constant strictly greater than 1. More importantly, (6.54) above shows that our method does not depend on the underlying dimension since (6.54) holds for any $d \geq 2$.

6.4 From Gaussian width bounds to sharp oracle inequalities

In this section, we develop yet another technique to derive sharp oracle inequalities for Least Squares estimators over closed convex sets. This technique is associated with localized Gaussian widths rather than statistical dimensions of cones considered above. The result is given in Theorem 6.12 below. Recently, other general methods have been proposed [27, 86, 103], but these methods did not provide oracle inequalities with leading constant 1.

Theorem 6.12. *Let $\mathcal{C}$ be a closed convex subset of $\mathbf{R}^n$, $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. Denote by $\Pi_{\mathcal{C}}(\boldsymbol{\mu})$ the projection of $\boldsymbol{\mu}$ onto $\mathcal{C}$. Assume that for some $t_* > 0$,*

$$\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{C}:\ |\Pi_{\mathcal{C}}(\boldsymbol{\mu})-\boldsymbol{u}|_2\le t_*} \boldsymbol{\xi}^T\left(\boldsymbol{u} - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\right)\right] \le \frac{t_*^2}{2}. \tag{6.55}$$

Then for any $x > 0$, with probability greater than $1 - e^{-x}$,

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}\|^2 \le \|\Pi_{\mathcal{C}}(\boldsymbol{\mu}) - \boldsymbol{\mu}\|^2 + \frac{2\max(t_*^2, 8\sigma^2 x)}{n}. \tag{6.56}$$

The proof of Theorem 6.12 is given in Section 6.7.3.

Note that condition (6.55) depends on the true vector $\boldsymbol{\mu}$ only through $\Pi_{\mathcal{C}}(\boldsymbol{\mu})$. The left hand side of (6.55) is the Gaussian width of $\mathcal{C}$ localized around $\Pi_{\mathcal{C}}(\boldsymbol{\mu})$. This differs from the recent analysis in [29] where the Gaussian width localized around $\boldsymbol{\mu}$ is studied. An advantage of considering the Gaussian width localized around $\Pi_{\mathcal{C}}(\boldsymbol{\mu})$ is that the resulting oracle inequality (6.56) is sharp. Chatterjee [29] proved that the Gaussian width localized around $\boldsymbol{\mu}$ characterizes a deterministic quantity $t_{\boldsymbol{\mu}}$ such that $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}\|$ concentrates around $t_{\boldsymbol{\mu}}\sqrt{n}$. This result from [29] grants both an upper bound and a lower bound on $|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}|_2$, but it does not imply nor is implied by a sharp oracle inequality such as (6.56) above. Thus, the result of [29] is of a different nature than (6.56).

Let $\mathcal{C}$ be a closed convex subset of $\mathbb{R}^n$. If the problem is misspecified, i.e., the true vector $\boldsymbol{\mu}$ does not belong to the set $\mathcal{C}$, two types of results arise naturally. Results of the first type are oracle inequalities such as (6.7) and (6.56) above. Results of the second type are upper bounds on the quantity

$$\|\hat{\boldsymbol{\mu}} - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\|^2, \tag{6.57}$$

if $\hat{\boldsymbol{\mu}}$ is an estimator such that $\hat{\boldsymbol{\mu}} \in \mathcal{C}$ almost surely. The quantity (6.57) is the estimation error of $\hat{\boldsymbol{\mu}}$ with respect to $\Pi_{\mathcal{C}}(\boldsymbol{\mu})$, the projection of $\boldsymbol{\mu}$ onto $\mathcal{C}$. The regret $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\Pi_{\mathcal{C}}(\boldsymbol{\mu}) - \boldsymbol{\mu}\|^2$ and the quantity (6.57) are two natural measures of the performance of $\hat{\boldsymbol{\mu}}$ under model misspecification. When $\hat{\boldsymbol{\mu}}$ is the Least Squares estimator over $\mathcal{C}$, (6.57) becomes $\|\Pi_{\mathcal{C}}(\boldsymbol{\mu}+\boldsymbol{\xi}) - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\|^2$. Estimation of $\Pi_{\mathcal{C}}(\boldsymbol{\mu})$ by the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C})$ has been considered in [107, Section 4] for $\mathcal{C} = \mathcal{S}_n^{\uparrow}$, and in [50, Section 6] for $\mathcal{C} = \mathcal{S}_n^{\cup}$. By convexity of $\mathcal{C}$ (cf. (6.11)), if $\hat{\boldsymbol{\mu}} \in \mathcal{C}$ then

$$\|\hat{\boldsymbol{\mu}} - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\|^2 \le \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\boldsymbol{\mu} - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\|^2. \tag{6.58}$$

Thus, sharp oracle inequalities such as (6.56) always imply upper bounds on $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\|^2$. On the other hand, oracle inequalities such as (6.7) with leading constant C strictly greater than 1 do not imply upper bounds on the estimation error (6.57).

A strategy to find a quantity t_* that satisfies (6.69) is to use metric entropy results together with Dudley integral bound, although Dudley integral bound may not be tight [20, Section 13.1, Exercises 13.4 and 13.5]. The following results are direct consequences of Theorem 6.12, Dudley integral bound and the entropy bounds from [45, 29, 50, 28].

Corollary 6.13. *There exists an absolute constant $C > 0$ such that the following holds. Let $n \geq 2$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$. Then for any $x > 0$, with probability greater than $1 - \exp(-x)$,*

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + C\left(\frac{D_{\boldsymbol{\mu}^*}\sigma^2}{n}\right)^{2/3} + \frac{16\sigma^2 x}{n}, \tag{6.59}$$

where $D_{\boldsymbol{\mu}^*} = \max(\sigma, \mu_n^* - \mu_1^*))$ and $\boldsymbol{\mu}^* = (\mu_1^*, ..., \mu_n^*)^T$ is the projection of $\boldsymbol{\mu}$ onto $\mathcal{S}_n^{\uparrow}$.

Corollary 6.14. *There exist absolute constants $\kappa, C > 0$ such that the following holds. Let $n \geq 3$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$ and that*

$$nR_{\boldsymbol{\mu}^*}^2 \geq \kappa \log(en)^{5/4}, \tag{6.60}$$

where $R_{\boldsymbol{\mu}^*} = \max(\sigma, \min(\{\|\boldsymbol{\mu}^* - \tau\|, \tau \in \mathbb{R}^n \text{ and } \tau \text{ is affine}\}))$ and $\boldsymbol{\mu}^*$ is the projection of $\boldsymbol{\mu}$ onto $\mathcal{S}_n^{\cup}$. Then for any $x > 0$, with probability greater than $1 - \exp(-x)$,

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{S}_n^{\cup}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{C\left(R_{\boldsymbol{\mu}^*}\sigma^4\right)^{2/5}\log(en)}{n^{4/5}} + \frac{16\sigma^2 x}{n}. \tag{6.61}$$

Corollary 6.15. *There exist absolute constants $C > 0$ such that the following holds. Let $d = 2$ and $n = n_1 n_2$ for two positive integers n_1, n_2. Let $\boldsymbol{\mu} \in \mathbf{R}^I$ where I is defined in (6.52), and let $\boldsymbol{\mu}^*$ be the projection of $\boldsymbol{\mu}$ onto the cone $\mathcal{K}^{2\uparrow}$ defined in (6.53). Then for all $x > 0$, with probability greater than $1 - \exp(-x)$,*

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}^{2\uparrow}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{K}^{2\uparrow}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{C\sigma^2\log(en)^8}{n} + \frac{C\sqrt{\sigma^2 V(\boldsymbol{\mu}^*)}}{n^{1/2}} + \frac{16\sigma^2 x}{n},$$

where $V(\boldsymbol{\mu}^*) = (1/n)\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}(\mu_{i_1 i_2}^* - \bar{\mu}^*)^2$ and $\bar{\mu}^* = (1/n)\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\mu_{i_1 i_2}^*$.

The novelty of Corollaries 6.13 to 6.15 are twofold. First, the leading constant is 1. Although model misspecification was considered in [107, 50], no oracle inequalities were obtained. Second, these sharp oracle inequalities hold in deviation, whereas the previous work derived upper bounds on the expected squared risk in the well-specified case. Note that one can derive sharp oracle inequalities in expectation by integrating the bounds of Corollaries 6.13, 6.14 and 6.15.

For any $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in \mathbb{R}^n$, let $|\boldsymbol{\mu}|_\infty = \max_{i=1,...,n} |\mu_i|$. It is easy to see that $\left|\Pi_{\mathcal{S}_n^{\uparrow}}(\boldsymbol{\mu})\right|_\infty \leq |\boldsymbol{\mu}|_\infty$. By integration, (6.59) implies the following bound on the maximal expected regret of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ over the class $\{\boldsymbol{\mu} \in \mathbb{R}^n : |\boldsymbol{\mu}|_\infty \leq D\}$

$$\sup_{\boldsymbol{\mu}\in\mathbb{R}^n : |\boldsymbol{\mu}|_\infty \leq D} \left(\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2\right) \leq C'\left(\frac{D\sigma^2}{n}\right)^{2/3}, \tag{6.62}$$

where $D \geq \sigma$ is a fixed parameter and $C' > 0$ is an absolute constant. Similar regret bounds hold for the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ with the rate $n^{-4/5}$, and for $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}^{2\uparrow})$ with the rate $n^{-1/2}$.

6.5 Aggregation of projections on opposite convex cones

Let $\mathcal{K}$ be a closed convex cone in $\mathbb{R}^n$. Let $-\mathcal{K} = \{-\boldsymbol{u} : \boldsymbol{u} \in \mathcal{K}\}$ be the opposite cone of $\mathcal{K}$. For instance, if $\mathcal{K} = \mathcal{S}_n^{\uparrow}$ is the cone of nondecreasing sequences, then $-\mathcal{K}$ is the cone of non-increasing sequences. If $\mathcal{K} = \mathcal{S}_n^{\cup}$ is the cone of convex sequences, then $-\mathcal{K}$ is the cone of concave sequences.

In this section we derive a sharp oracle inequality for the problem of aggregation of two estimators $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ and $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K})$. The goal is to construct an estimator $\hat{\boldsymbol{\mu}}$ that may depend on $\mathbf{y}$, $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ and $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K})$, such that with high probability,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \le \min_{\mathcal{C} \in \{\mathcal{K}, -\mathcal{K}\}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}\|^2 + \varepsilon,$$

where ε is a small quantity. Let us emphasize that no sample splitting is allowed, i.e., the same observation $\mathbf{y}$ is used to construct the estimators $\{\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}), \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K})\}$ and to perform the aggregation step. There exist procedures to aggregate with no sample splitting estimators of the form $A\mathbf{y}$ where A is a deterministic $n \times n$ matrix [70, 35, 33, 11], leading to sharp oracle inequalities. But to our knowledge there is no aggregation result of this type for nonlinear estimators. If $\mathcal{K}$ is not a subspace of $\mathbb{R}^n$, then the estimators $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ and $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K})$ are nonlinear estimators. Note that Theorem 6.16 substantially uses the fact that we have opposite cones. We do not know whether this can be extended to more general nonlinear estimators. Define the simplex in $\mathbf{R}^2$ by

$$\Lambda^2 = \{(\theta_+, \theta_-) \in \mathbf{R}^2, \quad \theta_+ \ge 0, \quad \theta_- \ge 0, \quad \theta_+ + \theta_- = 1\}.$$

For all $(\theta_+, \theta_-) \in \Lambda^2$, let

$$\hat{\boldsymbol{\mu}}_{(\theta_+,\theta_-)} = \theta_+ \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}) + \theta_- \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K}).$$

Finally, define the penalty

$$\mathrm{pen}(\theta_+, \theta_-) = \theta_+ \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}) - \hat{\boldsymbol{\mu}}_{(\theta_+,\theta_-)}\|^2 + \theta_- \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K}) - \hat{\boldsymbol{\mu}}_{(\theta_+,\theta_-)}\|^2.$$

We refer the reader to [11] for more details about this penalty.

Theorem 6.16. *Let $n \ge 2$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. Let $\mathcal{K} \subseteq \mathbb{R}^n$ be a closed convex cone. Let $\hat{\boldsymbol{\mu}}^*(\mathcal{K}) = \hat{\boldsymbol{\mu}}_{(\hat{\theta}_+,\hat{\theta}_-)}$ where*

$$(\hat{\theta}_+, \hat{\theta}_-) \in \operatorname*{argmin}_{(\theta_+,\theta_-) \in \Lambda^2} \|\mathbf{y} - \hat{\boldsymbol{\mu}}_{(\theta_+,\theta_-)}\|^2 + \frac{1}{2}\mathrm{pen}(\theta_+, \theta_-).$$

Then

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^*(\mathcal{K}) - \boldsymbol{\mu}\|^2 \le \min_{\mathcal{C} \in \{\mathcal{K}, -\mathcal{K}\}} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}\|^2 + \frac{4\sigma^2 \delta(\mathcal{K})}{n}. \tag{6.63}$$

Furthermore, for all $x > 0$, with probability greater than $1 - 2\exp(-x)$,

$$\|\hat{\boldsymbol{\mu}}^*(\mathcal{K}) - \boldsymbol{\mu}\|^2 \le \min_{\mathcal{C} \in \{\mathcal{K}, -\mathcal{K}\}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}\|^2 + \frac{4\sigma^2 \delta(\mathcal{K}) + 20\sigma^2 x}{n}, \tag{6.64}$$

where $\delta(\mathcal{K})$ is the statistical dimension (6.14) of the cone $\mathcal{K}$.

The proof of Theorem 6.16 is given in Section 6.7.4. The above aggregation procedure mimics the best estimator among the pair $\{\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}), \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K})\}$. The quantity $4\sigma^2\delta(\mathcal{K})/n$ may be referred to as the price to pay for aggregating the estimators $\{\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}), \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K})\}$. As $\mathbb{E}_{\boldsymbol{\mu}}\|\Pi_{\mathcal{K}}(\boldsymbol{\xi})\|^2 = \sigma^2\delta(\mathcal{K})/n$, this price is of order of the risk of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ when the true vector $\boldsymbol{\mu}$ is $\mathbf{0}$. More precisely, (6.63) can be rewritten as

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^*(\mathcal{K}) - \boldsymbol{\mu}\|^2 \le \min_{\mathcal{C}\in\{\mathcal{K},-\mathcal{K}\}} \mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}\|^2 + 4\mathbb{E}_{\mathbf{0}}\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K}) - \mathbf{0}\|^2.$$

To illustrate Theorem 6.16, let $\mathcal{K} = \mathcal{S}_n^{\uparrow}$ be the cone of nondecreasing sequences. By (6.20), $\delta(\mathcal{K}) \le \log(en)$. Let $k \in \{1, ..., n\}$ and consider the class $\mathcal{S}_n^{\uparrow\downarrow}(k) = \mathcal{S}_n^{[1]}(k) \cup (-\mathcal{S}_n^{[1]}(k))$, where the class $\mathcal{S}_n^{[1]}(k)$ is defined in (6.48). The class $\mathcal{S}_n^{\uparrow\downarrow}(k)$ is the set of all sequences that are either nondecreasing or non-increasing, and that are piecewise constant with less that k pieces. Combining Theorem 6.2 and the aggregation result above, we obtain the following bound on the maximal expected regret of $\hat{\boldsymbol{\mu}}^*(\mathcal{S}_n^{\uparrow})$

$$\sup_{\boldsymbol{\mu}\in\mathbb{R}^n}\left[\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^*(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u}\in\mathcal{S}_n^{\uparrow\downarrow}(k)}\|\boldsymbol{u} - \boldsymbol{\mu}\|^2\right] \le \frac{\sigma^2\Big(k\log(en/k) + 4\log(en)\Big)}{n}.$$

Similarly for higher order cones, combining Theorem 6.8 and the aggregation result above yields that for all integers s, β such that $1 \le \beta < n$ and $1 \le s \le n$,

$$\sup_{\boldsymbol{\mu}\in\mathbb{R}^n}\left[\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^*(\mathcal{S}_n^{[\beta]}) - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u}\in\Big(\mathcal{S}_n^{[\beta]}(s)\cup(-\mathcal{S}_n^{[\beta]}(s))\Big)}\|\boldsymbol{u} - \boldsymbol{\mu}\|^2\right]$$

is bounded from above by $\sigma^2 C(\beta)\Big(s\log(en/s) + 4\log(en)\Big)/n$ where $C(\beta)$ is the constant from (6.50). Thus the minimax regret for the class $\mathcal{S}_n^{[\beta]}(s) \cup (-\mathcal{S}_n^{[\beta]}(s))$ is of the order $\sigma^2 s/n$ (up to logarithmic factors), which is the order of the miminax regret for the class $\mathcal{S}_n^{[\beta]}(s)$.

6.6 Concluding remarks

We have presented two general methods to derive sharp oracle inequalities for the Least Squares estimator over a closed convex subset of $\mathbb{R}^n$. First, Proposition 6.3 shows that the Least Squares estimator over a closed convex set satisfies a sharp oracle inequality in deviation and expectation, where the remainder term is proportional to the sum of the statistical dimensions of some cones (cf. (6.36)). The second method is based on localized Gaussian widths and is given in Theorem 6.12. If $\mathcal{C}$ is a closed convex subset of $\mathbb{R}^n$, Theorem 6.12 shows that the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C})$ satisfies a sharp oracle inequality in deviation and expectation if the localized Gaussian width of $\mathcal{C}$ satisfies condition (6.55) for some constant $t_* > 0$. To summarize, our methods lead to the following improvements.

(i) Our oracle inequalities hold not only for the expected squared risk, but also in deviation with exponential probability tails. By integration, a sharp oracle inequality in deviation with exponential probability tails always implies a sharp oracle inequality in expectation. The reverse is not true, as there exist estimators that satisfy sharp oracle inequalities in expectation but not in deviation [3, 32].

(ii) Another improvement of the present paper over [107, 50, 27, 28] is that our oracle inequalities are sharp, i.e., with leading constant 1. Thus, our bounds account for model misspecification. This advantage can be interpreted at least in the following two ways.

(a) Let $\mathcal{C}$ be a closed convex set such that $\hat{\boldsymbol{\mu}} \notin \mathcal{C}$ and $\hat{\boldsymbol{\mu}}$ an estimator valued in $\mathcal{C}$. The quantity $\|\hat{\boldsymbol{\mu}} - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\|^2$ is a natural measure of the performance of $\hat{\boldsymbol{\mu}}$. As seen in (6.58), sharp oracle inequalities grant upper bounds on the quantity $\|\hat{\boldsymbol{\mu}} - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\|^2$, whereas oracle inequalities with leading constant strictly greater than 1 do not.

(b) A second advantage of sharp oracle inequalities is that they allow to bound from above the minimax regret. To see this, let $E, \bar{E}$ be two subsets of $\mathbb{R}^n$ with $E \subset \bar{E}$. If $\boldsymbol{\mu} \in E$, we say that the model is well-specified, if $\boldsymbol{\mu} \in \bar{E} \setminus E$ the model is misspecified. The minimax risk over E is $\inf_{\tilde{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in E} \mathbb{E}_{\boldsymbol{\mu}} \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ and the minimax regret with respect to $(E, \bar{E})$ is

$$\inf_{\tilde{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in \bar{E}} \left(\mathbb{E}_{\boldsymbol{\mu}} \|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \inf_{\boldsymbol{u} \in E} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \right),$$

where the infima are taken over all estimators. The minimax risk is a measure of the statistical complexity of E if the model is well-specified. The minimax regret is a natural measure of the statistical complexity of E that accounts for misspecification with respect to the set $\bar{E}$. There are situations where the minimax regret is substantially greater than the minimax risk [88]. Thus, it is important to study both the minimax risk and the minimax regret. As follows from (6.49) and (6.62), the sharp oracle inequalities (6.46) and (6.59) yield upper bounds on the minimax regret for $(E, \bar{E}) = (\mathcal{S}_n^{[\beta]}(s), \mathbb{R}^n)$ and $(E, \bar{E}) = (\mathcal{S}_n^{\uparrow}, \{\boldsymbol{\mu} \in \mathbb{R}^n : |\boldsymbol{\mu}|_\infty \leq D\})$. On the other hand, risk bounds or oracle inequalities with leading constant strictly greater than 1 do not imply bounds on the minimax regret.

6.7 Proofs

6.7.1 Upper bounds on statistical dimensions of cones

For any $k = 1, ..., n-1$, let $S_k = \{1, ..., k\}$ and $T_k = \{k+1, ..., n\}$. For any subset $T \subset \{1, ..., n\}$ and any vector $\boldsymbol{\theta} \in \mathbb{R}^n$, denote by $\boldsymbol{\theta}_T \in \mathbf{R}^T$ the restriction of $\boldsymbol{\theta}$ to T. For any $\boldsymbol{g}, \boldsymbol{\theta} \in \mathbb{R}^n$ and any $k = 1, ..., n-1$,

$$\boldsymbol{g}^T \boldsymbol{\theta} = \boldsymbol{g}_{T_k}^T \boldsymbol{\theta}_{T_k} + \boldsymbol{g}_{S_k}^T \boldsymbol{\theta}_{S_k}.$$

For any closed convex cone $\mathcal{K}$, denote by $\Pi_{\mathcal{K}}$ the projection onto $\mathcal{K}$.

Lemma 6.17. *Let $\mathcal{K} \subset \mathbb{R}^n$ be a closed convex cone. Assume that there exists a collection $\{(C_k^L, C_{n-k}^R), k = 1, ..., n-1\}$ where $C_k^L \subset \mathbb{R}^k, C_{n-k}^R \subset \mathbb{R}^{n-k}$ are closed convex cones such that the following holds. For all $\boldsymbol{\theta} \in \mathcal{K}$, there exists $k \in \{1, ..., n-1\}$ such that*

$$\boldsymbol{\theta}_{S_k} \in C_k^L, \qquad \boldsymbol{\theta}_{T_k} \in C_{n-k}^R. \tag{6.65}$$

Then

$$\delta(\mathcal{K}) \leq 2d^* + 6\log(n-1), \qquad \text{where} \quad d^* := \max_{k=1,...,n-1} \left[\delta(C_k^L) + \delta(C_{n-k}^R) \right].$$

Proof. Let $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$. If $\boldsymbol{\theta} \in \mathcal{K}$ is such that (6.65) holds for some $k = 1, ..., n-1$, then by the Cauchy-Schwarz inequality,

$$\boldsymbol{\theta}^T \boldsymbol{g} = \boldsymbol{\theta}_{S_k}^T \boldsymbol{g}_{S_k} + \boldsymbol{\theta}_{T_k}^T \boldsymbol{g}_{T_k} \le \sqrt{|\boldsymbol{\theta}_{S_k}^T|_2^2 + |\boldsymbol{\theta}_{T_k}^T|_2^2}\sqrt{Z_k^L + Z_{n-k}^R} = |\boldsymbol{\theta}|_2\sqrt{Z_k^L + Z_{n-k}^R}, \quad (6.66)$$

where we used the notation

$$Z_k^L := \left(\sup_{\boldsymbol{u}\in C_k^L : |\boldsymbol{u}|_2^2 \le 1} \boldsymbol{g}_{S_k}^T \boldsymbol{u}\right)^2, \qquad Z_{n-k}^R := \left(\sup_{\boldsymbol{u}\in C_{n-k}^R : |\boldsymbol{u}|_2^2 \le 1} \boldsymbol{g}_{T_k}^T \boldsymbol{u}\right)^2.$$

By (6.13), we have almost surely

$$Z_k^L = |\Pi_{C_k^L}(\boldsymbol{g}_{S_k})|_2^2, \qquad Z_{n-k}^R = |\Pi_{C_{n-k}^R}(\boldsymbol{g}_{T_k})|_2^2.$$

Similarly, let $Z := |\Pi_{\mathcal{K}}(\boldsymbol{g})|_2^2 = (\sup_{\boldsymbol{\theta}\in\mathcal{K}:|\boldsymbol{\theta}|_2^2\le 1} \boldsymbol{\theta}^T\boldsymbol{g})^2$. Using (6.65) and by taking the supremum over $|\boldsymbol{\theta}|_2^2 \le 1$ in (6.66) , we have established

$$\sqrt{Z} = \sup_{\boldsymbol{\theta}\in\mathcal{K}:|\boldsymbol{\theta}|_2^2\le 1} \boldsymbol{\theta}^T\boldsymbol{g} \le \max_{k=1,...,n-1} \sqrt{Z_k^L + Z_{n-k}^R}.$$

For any fixed k, $\boldsymbol{g}_{S_k}$ and $\boldsymbol{g}_{T_k}$ are independent, thus Z_k^L and Z_{n-k}^R are independent. By independence, for all $\lambda > 0$,

$$\mathbb{E}e^{\lambda Z} \le \sum_{k=1}^{n-1} \mathbb{E}\exp(\lambda Z_k^L + \lambda Z_{n-k}^R) = \sum_{k=1}^{n-1} \mathbb{E}[\exp(\lambda Z_k^L)]\,\mathbb{E}[\exp(\lambda Z_{n-k}^R)]. \quad (6.67)$$

Applying the moment generating function bound given in [1, Sublemma A.3], we obtain that for any $\lambda \in (0, 1/4)$,

$$\mathbb{E}[\exp(\lambda Z_k^L)]\,\mathbb{E}[\exp(\lambda Z_{n-k}^R)] \le \exp\left[\left(\frac{2\lambda^2}{1-4\lambda} + \lambda\right)\left(\delta\left(C_k^L\right) + \delta\left(C_{n-k}^R\right)\right)\right].$$

Let $\lambda^* = 1/6$. The previous display with $\lambda = \lambda^*$ and the definition of d^* yield

$$\mathbb{E}[\exp(\lambda^* Z_k^L)]\,\mathbb{E}[\exp(\lambda^* Z_{n-k}^R)] \le \exp\left[2\lambda^* d^*\right],$$

for all $k = 1, ..., n-1$. Plugging these bounds back into (6.67), we obtain

$$\mathbb{E}e^{\lambda^* Z} \le (n-1)e^{2\lambda^* d^*} = e^{\lambda^*[2d^* + 6\log(n-1)]}.$$

The function $t \to \exp(\lambda^* t)$ is convex, so applying Jensen's inequality yields $\delta(\mathcal{K}) = \mathbb{E}[Z] \le 2d^* + 6\log(n-1)$. □

Proof of Theorem 6.5. Define the convex cones

$$C_k^L = \mathcal{S}_k^{\downarrow}, \qquad C_{n-k}^R = \mathcal{S}_{n-k}^{\uparrow}, \qquad k = 1, ..., n-1, \quad (6.68)$$

i.e., the cone of non-increasing sequences in $\mathbb{R}^k$ and the cone of nondecreasing sequences in $\mathbb{R}^{n-k}$. A convex sequence $\boldsymbol{\theta} \in \mathcal{K}_{x_1,...,x_n}^C$ must be first non-increasing and then nondecreasing, so (6.65) holds for some $k = 1, ..., n-1$.

We apply Lemma 6.17 with $\mathcal{K} = \mathcal{K}_{x_1,...,x_n}^C$ and the cones defined in (6.68). By (6.20), $d^* \le 2\log(en)$, so Lemma 6.17 yields the bound $\delta(\mathcal{K}) \le 10\log(en)$. □

Proof of Theorem 6.7. We proceed by induction. Let $\beta \geq 1$. Assume that (6.45) holds for this β. We now prove that (6.45) holds for $\beta+1$. Define the convex cones

$$C_k^L = -\mathcal{S}_k^{[\beta]} \quad \text{if } k \geq \beta+1, \qquad C_{n-k}^L = \mathbb{R}^k \text{ otherwise,}$$
$$C_{n-k}^R = \mathcal{S}_{n-k}^{[\beta]}, \quad \text{if } n-k \geq \beta+1, \qquad C_k^R = \mathbb{R}^{n-k} \text{ otherwise,}$$

for all $k = 1, ..., n-1$, where $\mathcal{S}_k^{[\beta]} \subset \mathbb{R}^k$ and $\mathcal{S}_{n-k}^{[\beta]} \subset \mathbb{R}^{n-k}$ are defined in Example 6.3. Let $D_{n-\beta}, D_{n-\beta+1}, ..., D_n$ be the rectangular matrices defined in (6.17). Let $\boldsymbol{\theta} \in \mathcal{S}_n^{[\beta+1]}$ and define $\boldsymbol{v} = D_{n-\beta+1}...D_n\boldsymbol{\theta}$. As $\boldsymbol{\theta} \in \mathcal{S}_n^{[\beta+1]}$, $D_{n-\beta}\boldsymbol{v} \geq 0$. Thus $\boldsymbol{v}$ is a nondecreasing sequence in $\mathbb{R}^{n-\beta}$.

Let $E = \{l = 1, ..., n-\beta : v_l \leq 0\}$ where $v_l, l = 1, ..., n-\beta$ are the components of $\boldsymbol{v}$. If E is not empty, let $k = \beta - 1 + \max E$. Then $D_{k-\beta+1}...D_{k-1}D_k\boldsymbol{\theta}_{S_k} \leq 0$ so that $\boldsymbol{\theta}_{S_k} \in -\mathcal{S}_k^{[\beta]}$, and $D_{n-k-\beta+1}...D_{n-k-1}D_{n-k}\boldsymbol{\theta}_{T_k} \geq 0$ so that $\boldsymbol{\theta}_{T_k} \in C_{n-k}^R$. If E is empty, $\boldsymbol{v} > \boldsymbol{0}$ and thus $\boldsymbol{\theta} \in \mathcal{S}_n^{[\beta]}$, so (6.65) holds for $k = 1$. In summary, we have proved that for all $\boldsymbol{\theta} \in \mathcal{S}_n^{[\beta+1]}$, there exists $k = 1, ..., n-1$ such that (6.65) holds.

Combining Lemma 6.17 with (6.45) yields

$$\begin{aligned}\delta(\mathcal{S}_n^{[\beta+1]}) &\leq 4C(\beta)\log(en) + 6\log(n-1),\\ &\leq (4C(\beta)+6)\log(en) = C(\beta+1)\log(en),\end{aligned}$$

as by definition of $C(\beta+1)$ and $C(\beta)$, $4C(\beta)+6 = C(\beta+1)$. □

6.7.2 Lower bound

Define the support of $\boldsymbol{v} = (v_1, ..., v_n)^T \in \mathbb{R}^n$ by $\text{supp}(\boldsymbol{v}) = \{i = 1, ..., n : v_i \neq 0\}$.

Proof of Theorem 6.9. First, assume that $s \geq 9\beta + 1$. Let $S \geq 8$ be the largest integer such that $(S+1)\beta + 1 \leq s$. For all $\boldsymbol{\omega} = (\omega_1, ..., \omega_S)^T \in \{0,1\}^S$ and $\boldsymbol{\omega}' = (\omega_1', ..., \omega_S')^T \in \{0,1\}^S$, define the Hamming distance between $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ by $d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') = \sum_{k=1}^S |\omega_k - \omega_k'|$. By the Varshamov-Gilbert bound [99, Lemma 2.9], there exists $\Omega \subseteq \{0,1\}^S$ such that

$$\boldsymbol{0} = (0, ..., 0)^T \in \Omega, \quad \log(|\Omega| - 1) \geq S/8, \quad \text{and} \quad d_H(\boldsymbol{\omega}, \boldsymbol{\omega}') > S/8$$

for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$ such that $\boldsymbol{\omega} \neq \boldsymbol{\omega}'$. Let m be the largest integer such that $mS \leq n$. For each $\boldsymbol{\omega} \in \Omega$, define $\boldsymbol{u}^{\boldsymbol{\omega}} = (u_1^{\boldsymbol{\omega}}, ..., u_n^{\boldsymbol{\omega}})$ by

$$u_i^{\boldsymbol{\omega}} = \omega_j \qquad \text{if } jm \leq i-1 < jm+1, \qquad \text{for } i = 1, ..., Sm \text{ and } j = 1, ..., S,$$

and $u_i^{\boldsymbol{\omega}} = 0$ if $i > Sm$. Let $\Delta = \mathbb{X}^{-1}$ where $\mathbb{X}$ is the matrix (6.18), i.e., $\Delta\boldsymbol{u} = (u_1, u_2 - u_1, ..., u_n - u_{n-1})^T$ for all $\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n$. For all $\boldsymbol{\omega} \in \Omega$, $\boldsymbol{u}^{\boldsymbol{\omega}}$ is piecewise constant with at most $S+1$ pieces. It is easy to see that $\Delta\boldsymbol{u}^{\boldsymbol{\omega}}$ has at most $S+1$ nonzero components and

$$\text{supp}(\Delta\boldsymbol{u}^{\boldsymbol{\omega}}) \subset \cup_{k=0}^S \{km+1\}.$$

An immediate recurrence yields that for all $\boldsymbol{\omega} \in \Omega$,

$$\text{supp}(\Delta^\beta\boldsymbol{u}^{\boldsymbol{\omega}}) \subset \cup_{k=0}^S \{km+1, km+2, ..., km+\beta\}.$$

Let $T := (\cup_{k=0}^{S}\{km+1, km+2, ..., km+\beta\}) \cap \{1, ..., n\}$. We have $|T| \le (S+1)\beta$ and $\mathrm{supp}(\Delta^\beta \boldsymbol{u}^{\boldsymbol{\omega}}) \subset T$ for all $\boldsymbol{\omega} \in \Omega$. Define $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T \in \mathbb{R}^n$ by

$$\theta_i = \max\left(0, \max_{\boldsymbol{\omega}\in\Omega}\left(-(\Delta^\beta \boldsymbol{u}^{\boldsymbol{\omega}})_i\right)\right) \quad \text{if } i \in T, \qquad \theta_i = 0 \quad \text{if } i \notin T.$$

By construction, for all $\boldsymbol{\omega} \in \Omega$, $\mathrm{supp}(\boldsymbol{\theta} + \Delta^\beta \boldsymbol{u}^{\boldsymbol{\omega}}) \subset T$ and $\boldsymbol{\theta} + \Delta^\beta \boldsymbol{u}^{\boldsymbol{\omega}}$ has nonnegative entries. Let $\gamma = (1/8)\sqrt{\sigma^2/m}$. Using Lemma 6.18 below, for all $\boldsymbol{\omega} \in \Omega$,

$$\boldsymbol{x}^{\boldsymbol{\omega}} := \gamma \mathbb{X}^\beta\left(\boldsymbol{\theta} + \Delta^\beta \boldsymbol{u}^{\boldsymbol{\omega}}\right) = \gamma \mathbb{X}^\beta \boldsymbol{\theta} + \gamma \boldsymbol{u}^{\boldsymbol{\omega}}$$

belongs to $\mathcal{S}_n^{[\beta]}$, and $s_\beta(\boldsymbol{x}^{\boldsymbol{\omega}}) \le |T| + 1 \le \beta(S+1) + 1 \le s$. For two distinct $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$, we have

$$\|\boldsymbol{x}^{\boldsymbol{\omega}} - \boldsymbol{x}^{\boldsymbol{\omega}'}\|^2 = \frac{\gamma^2 d_H(\boldsymbol{\omega}, \boldsymbol{\omega}')m}{n} \ge \frac{\gamma^2 Sm}{8n} \ge \frac{\gamma^2}{16},$$

as by definition of m, $n/(2S) < m \le n/S$. Denote by $P_{\boldsymbol{\omega}}$ the distribution of $\boldsymbol{x}^{\boldsymbol{\omega}} + \boldsymbol{\xi}$. For any $\boldsymbol{\omega} \in \Omega$, the Kullback-Leibler divergence between the measures $P_{\boldsymbol{\omega}}$ and $P_{\mathbf{0}}$ satisfies

$$K(P_{\boldsymbol{\omega}}, P_{\mathbf{0}}) = \frac{n}{2\sigma^2}\|\boldsymbol{x}^{\boldsymbol{\omega}} - \boldsymbol{x}^{\mathbf{0}}\|^2 = \frac{n\gamma^2}{2\sigma^2}\|\boldsymbol{u}^{\boldsymbol{\omega}} - \boldsymbol{u}^{\mathbf{0}}\|^2 \le \frac{\gamma^2 Sm}{2\sigma^2} \le \frac{S}{128} \le \frac{\log|\Omega| - 1}{16}.$$

By [99, Theorem 2.7] with $\alpha = 1/16$, there exists an absolute constants $c > 0$ such that

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu}\in\mathcal{S}_n^{[\beta]}:\, s_\beta(\boldsymbol{\mu})\le s} \mathbb{P}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \ge \frac{\gamma^2}{64}\right) \ge c.$$

By definition of S we have $s \le 2S\beta$, and by definition of m we have $Sm \le n$. This implies that

$$64\gamma^2 = \frac{\sigma^2}{m} \ge \frac{\sigma^2 S}{n} \ge \frac{\sigma^2 s}{2\beta n}.$$

It remains to consider the case $s < 9\beta + 1$. In this case, the rate is of order $1/n$ and the lower bound follows from standard arguments by a reduction to testing between two simple hypotheses. □

Lemma 6.18. *Let $n < \beta$ be positive integers and let $\mathbb{X}$ be the matrix (6.18). Assume that $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T \in \mathbb{R}^n$ has nonnegative entries, i.e., $\theta_i \ge 0, i = 1, ..., n$. Then $\mathbb{X}^\beta \boldsymbol{\theta} \in \mathcal{S}_n^{[\beta]}$ and $s_\beta(\mathbb{X}^\beta \boldsymbol{\theta}) \le |\,\mathrm{supp}(\boldsymbol{\theta})| + 1$.*

Proof. The claim $\mathbb{X}^\beta \boldsymbol{\theta} \in \mathcal{S}_n^{[\beta]}$ follows from

$$D_{n-\beta+1} D_{n-\beta+2} ... D_n \mathbb{X}^\beta \boldsymbol{\theta} = (\theta_{\beta+1}, ..., \theta_n)^T.$$

Let $t_1 = 1$. There exists $k \ge 1$ such that $\mathrm{supp}(\boldsymbol{\theta}) \cup \{t_1\} = \{t_1, ..., t_k\}$ with $t_1 < ... < t_k$ and $k \le |\,\mathrm{supp}(\boldsymbol{\theta})| + 1$. Let $t_{k+1} = n + 1$ and define a partition $(T_1, ..., T_k)$ of $\{1, ..., n\}$ by $T_j = \{t_j, ..., t_{j+1} - 1\}, j = 1, ..., k$. Let $\boldsymbol{u}_j = (\mathbb{X}^\beta \boldsymbol{\theta})_{T_j} \in \mathbb{R}^{|T_j|}$. Let $j = 1, ..., k$. If $|T_j| \le \beta$ then using interpolation polynomials, the vector $\boldsymbol{u}_j$ satisfies $(\boldsymbol{u}_j)_i = Q_j(i), i = 1, ..., |T_j|$ for some polynomial Q_j of degree at most $\beta - 1$. If $|T_j| > \beta$ then

$$D_{|T_j|-\beta+1} ... D_{|T_j|} \boldsymbol{u}_j = (\theta_{t_j+\beta}, ..., \theta_{t_{j+1}-1})^T.$$

By definition of $t_1, ..., t_k$ we have $(\theta_{t_j+\beta}, ..., \theta_{t_{j+1}-1})^T = \mathbf{0}$. Thus $\boldsymbol{u}_j \in \mathrm{Lin}(\mathcal{S}_{|T_j|}^{[\beta]})$ and there exists a polynomial Q_j of degree at most $\beta - 1$ such that $(\boldsymbol{u}_j)_i = Q_j(i), i = 1, ..., |T_j|$. We have established that $s_\beta(\mathbb{X}^\beta \boldsymbol{\theta}) \le k \le |\,\mathrm{supp}(\boldsymbol{\theta})| + 1$. □

6.7.3 From Gaussian width bounds to sharp oracle inequalities

The proof of Theorem 6.12 is related to the isomorphic method [9] and the theory of local Rademacher complexities in regression with random design [10, 60].

Lemma 6.19. *Let $K \subseteq \mathbb{R}^n$. Assume that K is star-shaped at $\mathbf{0}$, i.e., that $\alpha \boldsymbol{u} \in K$ for all $\boldsymbol{u} \in K$ and all $\alpha \in [0,1]$. Let $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$. Assume that there exists $t_* > 0$ such that*

$$\mathbb{E}\left[\sup_{\boldsymbol{u}\in K:\, |\boldsymbol{u}|_2 \le t_*} \boldsymbol{\xi}^T \boldsymbol{u}\right] \le \frac{t_*^2}{2}. \tag{6.69}$$

Then for all $x > 0$, with probability greater than $1 - e^{-x}$ we have

$$2\boldsymbol{\xi}^T \boldsymbol{u} - |\boldsymbol{u}|_2^2 \le 2\max(t_*^2, 8\sigma^2 x) \tag{6.70}$$

simultaneously for all $\boldsymbol{u} \in K$.

Proof of Lemma 6.19. Let $\rho := \max(t_*, \sigma 2\sqrt{2x})$. The concentration inequality for suprema of Gaussian processes [20, Theorem 5.8] yields that on an event $\Omega(x)$ of probability greater than $1 - e^{-x}$,

$$Z := \sup_{\boldsymbol{u}\in K:\, |\boldsymbol{u}|_2 \le \rho} \boldsymbol{\xi}^T \boldsymbol{u} \le \mathbb{E}[Z] + \rho\sigma\sqrt{2x} \le \mathbb{E}[Z] + \frac{\rho^2}{2}.$$

Because K is star-shaped at $\mathbf{0}$, the function

$$\varphi : \quad t \to \frac{1}{t^2}\, \mathbb{E} \sup_{\boldsymbol{u}\in K:\, |\boldsymbol{u}|_2 \le t} \boldsymbol{\xi}^T \boldsymbol{u}$$

is non-increasing on $(0, +\infty)$. Indeed, for all $t > s > 0$, consider $\boldsymbol{u} \in K$ such that $|\boldsymbol{u}|_2 \le t$. Then $|(s/t)\boldsymbol{u}|_2 \le s$ and $(s/t)\boldsymbol{u} \in K$ because K is star-shaped at $\mathbf{0}$, hence

$$\frac{1}{t^2}\boldsymbol{\xi}^T \boldsymbol{u} = \frac{1}{st}\boldsymbol{\xi}^T((s/t)\boldsymbol{u}) \le \frac{1}{st} \sup_{\boldsymbol{v}\in K:\, |\boldsymbol{v}|_2 \le s} \boldsymbol{\xi}^T \boldsymbol{v} \le \frac{1}{s^2} \sup_{\boldsymbol{v}\in K:\, |\boldsymbol{v}|_2 \le s} \boldsymbol{\xi}^T \boldsymbol{v}.$$

Taking expectations yields that $\varphi(t) \le \varphi(s)$. Thus, the inequality $\rho \ge t_*$ and (6.69) imply that $\varphi(\rho) \le 1/2$ and $\mathbb{E}Z \le \rho^2/2$. Consequently, we have $Z \le \rho^2$ on $\Omega(x)$. Thus for all $\boldsymbol{u} \in K$ such that $|\boldsymbol{u}|_2 \le \rho$ we have $\boldsymbol{\xi}^T \boldsymbol{u} \le \rho^2$ and (6.70) holds.

If $\boldsymbol{u} \in K$ and $|\boldsymbol{u}|_2 > \rho$, then $\boldsymbol{\theta} = (\rho/|\boldsymbol{u}|_2)\boldsymbol{u}$ satisfies $|\boldsymbol{\theta}|_2 \le \rho$. By the above argument, we have $\boldsymbol{\xi}^T \boldsymbol{\theta} \le \rho^2$ on $\Omega(x)$. By rearranging, we get

$$2\boldsymbol{u}^T \boldsymbol{\xi} \le 2\rho|\boldsymbol{u}|_2 \le \rho^2 + |\boldsymbol{u}|_2^2 \le 2\rho^2 + |\boldsymbol{u}|_2^2,$$

which completes the proof of (6.70) for every $\boldsymbol{u} \in K$. □

Proof of Theorem 6.12. Inequality (6.16) with $\boldsymbol{u} = \Pi_{\mathcal{C}}(\boldsymbol{\mu})$ can be rewritten as

$$|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}|_2^2 - |\Pi_{\mathcal{C}}(\boldsymbol{\mu}) - \boldsymbol{\mu}|_2^2 \le 2\boldsymbol{\xi}^T \left(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \Pi_{\mathcal{C}}(\boldsymbol{\mu})\right) - |\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \Pi_{\mathcal{C}}(\boldsymbol{\mu})|_2^2.$$

Then, the claim is a direct consequence of Lemma 6.19 and (6.70) applied to $\boldsymbol{u} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \Pi_{\mathcal{C}}(\boldsymbol{\mu})$. □

Proof of Corollary 6.13. By rescaling we may assume that $\sigma = 1$. It was proved in [29, Inequality (51)] that there exists an absolute constant $c > 0$ such that $t_* := cD_{\boldsymbol{\mu}^*}1/3n^{1/6}$ satisfies (6.55). Applying Theorem 6.12 completes the proof. □

Proof of Corollary 6.14. By rescaling we may assume that $\sigma = 1$. Let $R = R_{\boldsymbol{\mu}^*}$. As in the proof of Corollary 6.13, we apply Theorem 6.12. We need to find $t_* > 0$ such that (6.69) holds for $K = \mathcal{S}_n^{\cup} - \boldsymbol{\mu}^*$. Let $r > 0$. Let $S(\boldsymbol{\mu}^*, r) = \{\boldsymbol{u} \in \mathcal{S}_n^{\cup}, \|\boldsymbol{u} - \boldsymbol{\mu}^*\| \le r\}$. By Dudley entropy bound (cf. [20, Corollary 13.2]), we obtain

$$\begin{aligned}\mathbb{E} \sup_{\boldsymbol{u} \in S(\boldsymbol{\mu}^*, r)} \frac{\boldsymbol{\xi}^T(\boldsymbol{u} - \boldsymbol{\mu}^*)}{\sqrt{n}} &\le 12 \int_0^r \sqrt{\log M(\varepsilon, S(\boldsymbol{\mu}^*, r), \|\cdot\|)} d\varepsilon, \\ &\le \bar{\kappa} \log(en)^{5/8} r^{3/4} (r^2 + R^2)^{1/8},\end{aligned}$$

where $\bar{\kappa} > 0$ is an absolute constant, $M(\varepsilon, S(\boldsymbol{\mu}^*, r), \|\cdot\|)$ is the ε-entropy of $S(\boldsymbol{\mu}^*, r)$ in the $\|\cdot\|$ norm, and the second inequality is proved in [50, (25)]. The constant $1/\sqrt{n}$ on the left hand side due to the fact that the Gaussian process is normalized with respect to the metric $\|\cdot\|$, i.e., for all vectors $\boldsymbol{u}, \boldsymbol{u}'$, $\|\boldsymbol{u} - \boldsymbol{u}'\|^2 = \mathbb{E}[(\boldsymbol{\xi}^T(\boldsymbol{u} - \boldsymbol{u}')/\sqrt{n})^2]$. Let now $t = r\sqrt{n}$. After rearranging, the previous inequality becomes

$$\mathbb{E} \sup_{\boldsymbol{u} \in \mathcal{S}_n^{\cup} : |\boldsymbol{\mu}^* - \boldsymbol{u}|_2^2 \le t} \boldsymbol{\xi}^T(\boldsymbol{u} - \boldsymbol{\mu}^*) \le \bar{\kappa} \log(en)^{5/8} n^{1/8} t^{3/4} \left(\frac{t^2}{n} + R^2\right)^{1/8}. \tag{6.71}$$

Let

$$t_* = \left(2\bar{\kappa} 2^{1/8}\right)^{4/5} \sqrt{\log(en)} R^{1/5} n^{1/10},$$

and choose the absolute constant $\kappa := 4\bar{\kappa}^2 2^{1/4}$. With this choice of κ and t_*, (6.60) is equivalent to $t_*^2/n \le R^2$. Thus, for $t = t_*$, the right hand side of (6.71) does not exceed

$$\bar{\kappa} 2^{1/8} \log(en)^{5/8} n^{1/8} t_*^{3/4} \le t_*^2/2.$$

Applying Theorem 6.12 completes the proof of (6.61). □

Proof of Corollary 6.15. It was proved in [28, Section 2] that there exists an absolute constant $c > 0$ such that $t_* := c\sqrt{\sigma^2 \log(en)^8/n + \sqrt{\sigma^2 V(\boldsymbol{\mu}^*)/n}}$ satisfies (6.55). Applying Theorem 6.12 completes the proof. □

6.7.4 Aggregation on opposite cones

Proof of Theorem 6.16. For notational simplicity, let $\hat{\boldsymbol{\mu}}^* = \hat{\boldsymbol{\mu}}^*(\mathcal{K})$, $\hat{\boldsymbol{\mu}}^+ = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{K})$ and $\hat{\boldsymbol{\mu}}^- = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(-\mathcal{K})$. Let $H(\theta_+, \theta-) = \|\mathbf{y} - \hat{\boldsymbol{\mu}}_{(\theta_+, \theta_-)}\|^2 + \frac{1}{2}\mathrm{pen}(\theta_+, \theta_-)$. The function H is convex and differentiable, thus it satisfies

$$\nabla H(\hat{\theta}_+, \hat{\theta}_-)^T \Big((1, 0) - (\hat{\theta}_+, \hat{\theta}_-)\Big) \ge 0,$$

cf. [21, 4.2.3, equation (4.21)]. This inequality can be rewritten as

$$\|\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}\|^2 - \|\hat{\boldsymbol{\mu}}^+ - \boldsymbol{\mu}\|^2 \le \frac{2}{n}\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}^* - \hat{\boldsymbol{\mu}}^+) - \frac{1}{2}\mathrm{pen}(\theta_+, \theta_-) - \frac{1}{2}\|\hat{\boldsymbol{\mu}}^+ - \hat{\boldsymbol{\mu}}^*\|^2.$$

Simple algebra yields that the right hand side of the previous display is equal to

$$\frac{2}{n}\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}^* - \hat{\boldsymbol{\mu}}^+) - \frac{1}{2}\hat{\theta}_-\|\hat{\boldsymbol{\mu}}^+ - \hat{\boldsymbol{\mu}}_-\|^2 =: L(\hat{\theta}_+, \hat{\theta}_-).$$

The function L is linear, thus it is maximized at a vertex of Λ^2, and

$$\|\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}\|^2 - \|\hat{\boldsymbol{\mu}}^+ - \boldsymbol{\mu}\|^2 \leq \max\left(0, \frac{2}{n}\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}^- - \hat{\boldsymbol{\mu}}^+) - \frac{1}{2}\|\hat{\boldsymbol{\mu}}^+ - \hat{\boldsymbol{\mu}}^-\|^2\right).$$

Note that $\hat{\boldsymbol{\mu}}^- - \hat{\boldsymbol{\mu}}^+ \in -\mathcal{K}$. Thus by definition of the projection on $-\mathcal{K}$ denoted by $\Pi_{-\mathcal{K}}$,

$$\begin{aligned}\frac{4}{n}\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}^- - \hat{\boldsymbol{\mu}}^+) - \|\hat{\boldsymbol{\mu}}^+ - \hat{\boldsymbol{\mu}}^-\|^2 &= \|2\boldsymbol{\xi}\|^2 - \|2\boldsymbol{\xi} - (\hat{\boldsymbol{\mu}}^- - \hat{\boldsymbol{\mu}}^+)\|^2, \\ &\leq \|2\boldsymbol{\xi}\|^2 - \|2\boldsymbol{\xi} - \Pi_{-\mathcal{K}}(2\boldsymbol{\xi})\|^2 = \|\Pi_{-\mathcal{K}}(2\boldsymbol{\xi})\|^2.\end{aligned}$$

In summary, we have proved that

$$\|\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}\|^2 - \|\hat{\boldsymbol{\mu}}^+ - \boldsymbol{\mu}\|^2 \leq 2\|\Pi_{-\mathcal{K}}(\boldsymbol{\xi})\|^2.$$

Similarly, $\nabla H(\hat{\theta}_+, \hat{\theta}_-)^T\left((0,1) - (\hat{\theta}_+, \hat{\theta}_-)\right) \geq 0$ implies

$$\|\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}\|^2 - \|\hat{\boldsymbol{\mu}}^- - \boldsymbol{\mu}\|^2 \leq 2\|\Pi_{\mathcal{K}}(\boldsymbol{\xi})\|^2.$$

Combining the two bounds, we obtain

$$\|\hat{\boldsymbol{\mu}}^*(\mathcal{K}) - \boldsymbol{\mu}\|^2 \leq \min_{\mathcal{C}\in\{\mathcal{K},-\mathcal{K}\}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{C}) - \boldsymbol{\mu}\|^2 + 2\max\left(\|\Pi_{\mathcal{K}}(\boldsymbol{\xi})\|^2, \|\Pi_{-\mathcal{K}}(\boldsymbol{\xi})\|^2\right).$$

Using that $\max(a,b) \leq a+b$ for all $a,b>0$ and taking the expectation of the previous display yields (6.63). To prove (6.64), we apply Lemma 6.20 with $t=x$, $k=1$ and $a=1$ twice, to the cones $\mathcal{K}$ and $-\mathcal{K}$. The union bound and the equality $\delta(\mathcal{K}) = \delta(-\mathcal{K})$ complete the proof. □

6.7.5 Concentration lemma

Lemma 6.20. *Let k be a positive integer. For all $j=1,...,k$, let d_j be a positive integer, let $\mathcal{K}_j$ be a closed convex cone in $\mathbf{R}^{d_j}$, and let Π_j be the projection onto $\mathcal{K}_j$. Let $\boldsymbol{\xi}_1,...,\boldsymbol{\xi}_k$ be independent random variables with $\boldsymbol{\xi}_j \sim \mathcal{N}(0, \sigma^2 I_{d_j\times d_j})$ for all $j=1,...,k$. Then for all $t>0$ and $a>0$, with probability greater than $1-\exp(-t)$,*

$$\frac{1}{\sigma^2}\sum_{j=1}^k |\Pi_j(\boldsymbol{\xi}_j)|_2^2 \leq (1+a)\left(\sum_{j=1}^k \delta_j\right) + \left(8 + \frac{2}{a}\right)t,$$

where for all $j=1,...,k$, $\delta_j = \mathbb{E}\left[|\Pi_j(\boldsymbol{\xi}_j)|_2^2\right]$.

The Lemma above is a consequence of the moment generating function bound given in [1, Sublemma E.3].

Proof. By homogeneity it is enough to prove the Lemma for $\sigma=1$. By the independence of $\boldsymbol{\xi}_1,...,\boldsymbol{\xi}_k$ and [1, Sublemma E.3], we get that for all $u \in (-1/4, 1/4)$,

$$\mathbb{E}\exp\left(-u\left(\sum_{j=1}^k\left(|\Pi_j(\boldsymbol{\xi}_j)|_2^2 - \delta_j\right)\right)\right) \leq \exp\frac{2u^2\sum_{j=1}^k\delta_j}{1-4|u|}.$$

Let $S = \sum_{j=1}^{k} \delta_j$. By Markov exponential inequality, for all $\lambda > 0$ and for $u = \lambda/(4(S+\lambda))$,

$$\mathbb{P}\left(\sum_{j=1}^{k} |\Pi_j(\boldsymbol{\xi}_j)|_2^2 \geq S + \lambda\right) \leq \exp\left(-\lambda u + \frac{2u^2 S}{1 - 4|u|}\right) = \exp\left(-\frac{\lambda^2/8}{S+\lambda}\right).$$

For all $t > 0$, we now set $\lambda = 2\sqrt{2St + 4t^2} + 4t$ so that $t = \frac{\lambda^2/8}{S+\lambda}$. To complete the proof, observe that for all $a > 0$,

$$\lambda \leq 2\sqrt{2St} + 8t \leq aS + (8 + 2/a)t.$$

□

Chapter 7

Adaptive confidence sets in shape restricted regression

We construct adaptive confidence sets in isotonic and convex regression. In univariate isotonic regression, if the true parameter is piecewise constant with k pieces, then the Least-Squares estimator achieves a parametric rate of order k/n up to logarithmic factors. We construct honest confidence sets that adapt to the unknown number of pieces of the true parameter. The proposed confidence set enjoys uniform coverage over all non-decreasing functions. Furthermore, the squared diameter of the confidence set is of order k/n up to logarithmic factors, which is optimal in a minimax sense. In univariate convex regression, we construct a confidence set that enjoys uniform coverage and such that its diameter is of order q/n up to logarithmic factors, where $q-1$ is the number of changes of slope of the true regression function.

7.1 Introduction

Let $K \subset \mathbb{R}^n$ be a closed convex set. Assume that we have the observations

$$Y_i = \mu_i + \xi_i, \qquad i = 1, ..., n,$$

where the vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T \in K$ is unknown, $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$ is a noise vector with n-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ where $\sigma > 0$ and $I_{n\times n}$ is the $n \times n$ identity matrix. Denote by $\mathbb{E}_{\boldsymbol{\mu}}$ and $\mathbb{P}_{\boldsymbol{\mu}}$ the expectation and the probability measure corresponding to the distribution of the random variable $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\xi}$. The vector $\mathbf{y} = (Y_1, ..., Y_n)^T$ is observed and the goal is to estimate $\boldsymbol{\mu}$. Consider the scaled norm $\|\cdot\|$ defined by

$$\|\boldsymbol{u}\|^2 = \frac{1}{n}\sum_{i=1}^{n} u_i^2, \qquad \boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n.$$

The error of an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ is given by $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$. Let $|\cdot|_2^2$ be the squared Euclidean norm, so that $\frac{1}{n}|\cdot|_2^2 = \|\cdot\|^2$. For a finite set E, let $|E|$ denote its cardinality. We use bold face for vectors and the components of any vector $\boldsymbol{v} \in \mathbb{R}^n$ are denoted by $v_1, ..., v_n$.

In this paper, we consider the particular case where K is a polyhedron, that is, an intersection of a finite number of half-spaces. If the true parameter $\boldsymbol{\mu}$ lies in a low-dimensional face of the polyhedron K, it has been shown that for some polyhedra K, the rate of estimation is of order $\frac{d\sigma^2}{n}$ up to logarithmic factors, where d is the

dimension of the smallest face that contains $\boldsymbol{\mu}$ [50, 27, 28, 13]. This phenomenon appears, for example, if the polyhedron K is the cone of nondecreasing sequences [27, 13] or the cone of convex sequences [50, 13]. For these examples, if $\boldsymbol{\mu}$ lies in a d-dimensional face of the polyhedron K, the Least Squares estimator over K satisfy risk bounds and oracle inequalities with the parametric rate $\frac{d\sigma^2}{n}$, up to logarithmic factors. We consider the problem of confidence sets in this context. In particular, the present paper addresses the following questions.

- Is it possible to estimate or bound from below by a data-driven quantity the dimension d of the smallest face of the polyhedron K that contains the true parameter $\boldsymbol{\mu}$?
- Is it possible to construct a confidence set $\hat{C}_n$ such that:
  1. It enjoys uniform coverage over all $\boldsymbol{\mu} \in K$ (i.e., $\boldsymbol{\mu} \in \hat{C}_n$ with high probability).
  2. It adapts to the smallest low-dimensional face that contains $\boldsymbol{\mu}$ (i.e., the diameter of $\hat{C}_n$ should be of the order $\frac{d\sigma^2}{n}$ up to logarithmic factors if the smallest face that contains $\boldsymbol{\mu}$ has dimension d).

In this paper, we answer these questions for two particular polyhedra: the cone of nondecreasing sequences and the cone of convex sequences.

The construction of adaptive confidence sets in isotonic or convex regression has been studied in [42, 22, 23]. These papers show that if the true regression function is simultaneously smooth and monotone, then it is possible to construct confidence sets that adapt to the unknown smoothness of the true regression function. In the present paper, there is no smoothness assumption and the goal is to construct confidence sets that adapt to the dimension d of the smallest face of the polyhedron.

The rest of the paper is organized as follows. Section 7.2 gives the definition of honest and adaptive confidence sets. Section 7.3 defines the cone of nondecreasing sequences and recalls some material from [1, 77] on the statistical dimension and the intrinsic volumes of closed convex cones. In Section 7.4 and Section 7.6, we construct honest and adaptive confidence sets for the cone of nondecreasing sequences and for the cone of convex sequences, respectively. We discuss the scope of these results in Section 7.7.

7.2 Honest and adaptive confidence sets

Let $(E_k)_{k\in J}$ be a collection of subsets K indexed by some possibly infinite set J. We will refer to the sets $(E_k)_{k\in J}$ as the *models*. If $J = \{1, ..., k_{\text{MAX}}\}$, these models may be ordered by inclusion so that

$$E_1 \subset ... \subset E_{k_{\text{MAX}}} = K. \tag{7.1}$$

For any model $E_k \subset K$, the minimax risk on E_k is the quantity

$$R^*_{\mathbb{E}}(E_k) = \inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu}\in E_k} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2,$$

where the infimum is taken over all estimators, that is, all random variables of the form $\hat{\boldsymbol{\mu}} = g(\mathbf{y})$ where $g : \mathbb{R}^n \to \mathbb{R}^n$ is a Borel function. If $J = \{1, ..., k_{\text{MAX}}\}$ and (7.1)

holds, the minimax risks satisfy

$$R^*_{\mathbb{E}}(E_1) \leq ... \leq R^*_{\mathbb{E}}(E_{k_{\text{MAX}}}).$$

In that case, the collection $(E_k)_{k=1,...,k_{\text{MAX}}}$ represents models of increasing complexity.

Similarly, if a confidence value $\alpha \in (0,1)$ is given, one may define the minimax quantity

$$R^*_\alpha(E_k) = \inf\left\{R > 0 : \quad \sup_{\hat{\boldsymbol{\mu}}} \inf_{\boldsymbol{\mu} \in E_k} \mathbb{P}_{\boldsymbol{\mu}}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq R\right) \geq 1 - \alpha\right\} \tag{7.2}$$

for all $k \in J$, where the supremum of taken over all estimators. This quantity represents the smallest size, if a minimax sense, of a confidence ball with confidence level $1-\alpha$. Similarly, if $J = \{1, ..., k\,\text{max}\}$ and the models are ordered by inclusion as in (7.1), this quantity is an increasing function of k and we have

$$R^*_\alpha(E_1) \leq ... \leq R^*_\alpha(E_{k_{\text{MAX}}}).$$

for all $\alpha \in (0,1)$.

The goal of this paper is to study confidence sets in shape restricted regression. A confidence set is a region $\hat{C}_n$ such that with high probability, the unknown parameter $\boldsymbol{\mu}$ belongs to $\hat{C}_n$. Let $\alpha \in (0,1)$. If $\boldsymbol{\mu} \in E_{k^*}$ for some $k^* \in J$, the quantity (7.2) may be used to define the oracle region

$$\hat{C}^*_n(k^*) := \{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u} - \hat{\boldsymbol{\mu}}\|^2 \leq R^*_\alpha(E_k)\},$$

where $\hat{\boldsymbol{\mu}}$ is an estimator that achieves the supremum in (7.2) (we assume here that all infima and suprema in (7.2) are attained). Then, by definition of $R^*_\alpha(\cdot)$, we have that $\boldsymbol{\mu} \in \hat{C}^*_n(k^*)$ with probability at least $1-\alpha$. We call $\hat{C}^*_n(k^*)$ an *oracle* region since it is inaccessible for two reasons.

First, the radius $R^*_\alpha(E_{k^*})$ and the integer k^* must be known in order to construct $\hat{C}^*_n(k^*)$, i.e., the knowledge of the smallest model that contains $\boldsymbol{\mu}$ is needed. Second, the oracle region $\hat{C}^*_n(k^*)$ is an Euclidean ball centered around the estimator $\hat{\boldsymbol{\mu}}$ that achieves the infimum in (7.2), and this estimator is unknown.

This paper studies the construction of data-driven confidence sets $\hat{C}_n$. We consider only $1-\alpha$ confidence sets, which means that the true parameter $\boldsymbol{\mu}$ belongs to $\hat{C}_n$ with probability at least $1-\alpha$, uniformly over all $\boldsymbol{\mu} \in K$.

We also want the diameter of the confidence set $\hat{C}_n$ to be of the same order as the diameter of the oracle region $\hat{C}^*_n(k^*)$, that is, the value $R^*_\alpha(E_{k^*})$. Furthermore the construction of $\hat{C}_n$ should not require the knowledge of the smallest model that contains the true parameter $\boldsymbol{\mu}$: The knowledge of k^* is not needed to construct the confidence region $\hat{C}_n$. In that case, we say that the confidence set $\hat{C}_n$ is adaptive.

We now give a formal definition of these properties. For any $A \subset \mathbb{R}^n$, define the diameter of A for the scaled norm $\|\cdot\|$ by

$$\text{diam}\, A := \sup_{\boldsymbol{v},\boldsymbol{u} \in A} \|\boldsymbol{v} - \boldsymbol{u}\|.$$

Definition 7.1. Let $\alpha \in (0,1)$. Let $K \subset \mathbb{R}^n$ be a closed convex set and let $(E_k)_{k \in J}$ be a collection of subsets of K indexed by an arbitrary set J. Let $\hat{C}_n = \hat{C}_n(\mathbf{y})$ be a Borel subset of $\mathbb{R}^n$ measurable with respect to $\mathbf{y}$. We say that $\hat{C}_n$ is an honest confidence set if

$$\inf_{\boldsymbol{\mu} \in K} \mathbb{P}_{\boldsymbol{\mu}}\left(\boldsymbol{\mu} \in \hat{C}_n\right) \geq 1 - \alpha. \tag{7.3}$$

We say that an honest confidence set $\hat{C}_n$ is adaptive in probability if for all $\gamma \in (0,1)$,

$$\inf_{k\in J}\inf_{\boldsymbol{\mu}\in E_k} \mathbb{P}_{\boldsymbol{\mu}}\left(\operatorname{diam}(\hat{C}_n)^2 \le c' R^*_\alpha(E_k)\log\left(\frac{en}{\gamma\alpha}\right)^c\right) \ge 1-\gamma, \tag{7.4}$$

where $c' > 0$ and $c \ge 0$ are numerical constants. Alternatively to (7.4), we say that the confidence set $\hat{C}_n$ is adaptive in expectation if for all $k \in J$,

$$\sup_{\boldsymbol{\mu}\in E_k} \mathbb{E}_{\boldsymbol{\mu}}\left[\operatorname{diam}(\hat{C}_n)^2\right] \le c' R^*_{\mathbb{E}}(E_k)\log\left(\frac{en}{\alpha}\right)^c, \tag{7.5}$$

where $c' > 0$ and $c \ge 0$ are numerical constants.

The role of the constant $c \ge 0$ is to allow for logarithmic factors.

The statistic $\hat{C}_n$ induces a confidence set. If the definition above holds, (7.3) says that the true sequence $\boldsymbol{\mu}$ lies in $\hat{C}_n$ with high probability. Inequality (7.4) implies that if the true parameter satisfies $\hat{\boldsymbol{\mu}} \in E_{k^*}$ for some $k^* \in J$, then the diameter of $\hat{C}_n$ is of the same order as the minimax quantity (7.2) of the model E_k, up to logarithmic factors.

We now consider a special case: confidence ball around the Least Squares estimator. The Least Squares estimator over a closed convex set K is defined by

$$\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) = \operatorname*{argmin}_{\boldsymbol{u}\in K} \|\mathbf{y}-\boldsymbol{u}\|^2 = \Pi_K(\mathbf{y})$$

where Π_K denotes the convex projection onto K. By definition of the convex projection onto K, we have $(\boldsymbol{u}-\Pi_K(\mathbf{y}))^T(\mathbf{y}-\Pi_K(\mathbf{y})) \le 0$ for all $\boldsymbol{u} \in K$, which can be rewritten as

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)-\mathbf{y}\|^2 \le \|\boldsymbol{u}-\mathbf{y}\|^2 - \|\boldsymbol{u}-\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)\|^2. \tag{7.6}$$

If the confidence set $\hat{C}_n$ is an Euclidean ball, it is characterized by its center and its radius.

Let $\alpha \in (0,1)$ be a confidence value. Let $K \subset \mathbb{R}^n$ be a closed convex set and let $(E_k)_{k\in J}$ be a collection of subsets of K indexed by an arbitrary set J. Let $\hat{r}$ be a positive random variable measurable with respect to $\mathbf{y}$ and let $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)$ be the Least Squares estimator over K. The set

$$\hat{C}_n = \{\boldsymbol{v} \in \mathbb{R}^n : \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)-\boldsymbol{v}\|^2 \le \hat{r}\} \tag{7.7}$$

is an honest confidence ball if (7.3) holds. The confidence ball $\hat{C}_n$ is said to be adaptive in probability if (7.4) holds, that is, for all $\gamma \in (0,1)$,

$$\inf_{k\in J}\inf_{\boldsymbol{\mu}\in E_k} \mathbb{P}_{\boldsymbol{\mu}}\left(\hat{r} \le c' R^*_\alpha(E_k)\log\left(\frac{en}{\gamma\alpha}\right)^c\right) \ge 1-\gamma, \tag{7.8}$$

for all $\gamma \in (0,1)$ where $c' > 0$ and $c \ge 0$ are numerical constants. The confidence ball $\hat{C}_n$ is said to be adaptive in expectation if (7.5), that is,

$$\sup_{\boldsymbol{\mu}\in E_k} \mathbb{E}_{\boldsymbol{\mu}}[\hat{r}] \le c' R^*_{\mathbb{E}}(E_k)\log\left(\frac{en}{\alpha}\right)^c, \tag{7.9}$$

for all $k \in J$, where $c' > 0$ and $c \ge 0$ are numerical constants.

7.3 Preliminaries

7.3.1 The cone of nondecreasing sequences and the models $\left(\mathcal{S}_n^{\uparrow}(k)\right)_{k=1,\dots,n}$

Let $\mathcal{S}_n^{\uparrow}$ be the set of all nondecreasing sequences, defined by

$$\mathcal{S}_n^{\uparrow} := \{\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathbb{R}^n : u_i \leq u_{i+1}, \quad i = 1, ..., n-1\}.$$

For $n = 1$, let $\mathcal{S}_n^{\uparrow} = \mathbb{R}$. For all $n \geq 1$, define the cone of non-increasing sequences by $\mathcal{S}_n^{\downarrow} := -\mathcal{S}_n^{\uparrow}$.

For any $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$, let $k(\boldsymbol{u}) := |\{u_i, i = 1, ..., n\}|$ where $|A|$ denotes the cardinality of set A. The integer $k(\boldsymbol{u})$ is the smallest positive integer such that $\boldsymbol{u}$ is piecewise constant with $k(\boldsymbol{u})$ pieces. The integer $k(\boldsymbol{u}) - 1$ is also the number of jumps of $\boldsymbol{u}$, that is, the number of inequalities $u_i \leq u_{i+1}$ that are strict. Define the sets

$$\mathcal{S}_n^{\uparrow}(k) = \{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow} : k(\boldsymbol{u}) \leq k\}, \qquad k = 1, ..., n.$$

The set $\mathcal{S}_n^{\uparrow}(1)$ is the subspace of all constant sequences while $\mathcal{S}_n^{\uparrow}(2), ..., \mathcal{S}_n^{\uparrow}(n-1)$ are closed non-convex sets. We have

$$\mathcal{S}_n^{\uparrow}(1) \subset \mathcal{S}_n^{\uparrow}(2) \subset ... \subset \mathcal{S}_n^{\uparrow}(n) = \mathcal{S}_n^{\uparrow}.$$

It is known that there exist numerical constants c, c' such that for all $\alpha \leq c$,

$$\frac{c'\sigma^2 k}{n} \leq R_{\alpha}^*(\mathcal{S}_n^{\uparrow}(k)) \leq \frac{2\sigma^2 k \log(en/k)}{n} + \frac{10 \log(1/\alpha)}{n},$$

cf. [14, Proposition 4] for the lower bound and [13] for the upper bound. Thus, for $\alpha > 0$ small enough, the quantity $R_{\alpha}^*(\mathcal{S}_n^{\uparrow}(k))$ is of order $k\sigma^2/n$, up to logarithmic factors in n and $1/\alpha$. Furthermore, the minimax risk over the sets $\mathcal{S}_n^{\uparrow}(k)$ satisfies

$$\frac{c''\sigma^2 k}{n} \leq R_{\mathbb{E}}^*(\mathcal{S}_n^{\uparrow}(k)) \leq \sup_{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}(k)} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \frac{\sigma^2 k \log(en/k)}{n}, \tag{7.10}$$

for some numerical constant $c'' > 0$, cf. [13, Theorem 2] for the upper bound and [14, (30)] for the lower bound. Finally, (7.10) implies that the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ achieves the minimax rate, up to logarithmic factors.

7.3.2 Statistical dimension and intrinsic volumes of cones

We recall here some properties of closed convex cones. Most of the material of the present section comes from [1, 77]. In the present paper, a cone is always pointed at 0. A polyhedral cone is a closed convex cone of the form

$$K = \{\boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{u}^T \boldsymbol{v}_j \leq 0 \text{ for all } j = 1, ..., k\}, \tag{7.11}$$

where $\boldsymbol{v}_1, ..., \boldsymbol{v}_k$ are vectors in $\mathbb{R}^n$, that is, K is the intersection of a finite number of half-spaces. The dual or polar cone of K is defined as

$$K^{\circ} := \{\boldsymbol{\theta} \in \mathbb{R}^n : \boldsymbol{v}^T \boldsymbol{\theta} \leq 0 \text{ for all } \boldsymbol{v} \in K\}.$$

If K a polyhedral cone, the face of K with outward vector $\boldsymbol{\theta} \in \mathbb{R}^n$ is the set

$$F(\boldsymbol{\theta}) := \{\boldsymbol{u} \in K : \boldsymbol{u}^T\boldsymbol{\theta} = \sup_{\boldsymbol{v}\in K} \boldsymbol{v}^T\boldsymbol{\theta}\}.$$

The face $F(\boldsymbol{\theta})$ is nonempty if and only if $\boldsymbol{\theta} \in K^\circ$. If K is the polyhedral cone (7.11) defined by the vectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_k$, a face of a polyhedral cone K has to be of the form

$$\{\boldsymbol{u} \in K : \boldsymbol{u}^T\boldsymbol{v}_j = 0 \text{ for all } j \in T\} \tag{7.12}$$

for some $T \subset \{1, ..., k\}$. The dimension of a face F is the dimension of the linear span of F.

Definition 7.2 (Statistical dimension, Amelunxen et al. [1])**.** For any closed convex cone $K \subset \mathbb{R}^n$, define

$$\delta(K) := \mathbb{E}\left[|\Pi_K(\boldsymbol{g})|_2^2\right] = \mathbb{E}\left[\boldsymbol{g}^T\Pi_K(\boldsymbol{g})\right] = \mathbb{E}\left[\left(\sup_{\boldsymbol{\theta}\in K:|\boldsymbol{\theta}|_2\le 1} \boldsymbol{g}^T\boldsymbol{\theta}\right)^2\right],$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, I_{n\times n})$. The quantity $\delta(K)$ is called the statistical dimension of the cone K.

It is also known that the following holds almost surely

$$|\Pi_K(\boldsymbol{g})|_2^2 = \boldsymbol{g}^T\Pi_K(\boldsymbol{g}) = \left(\sup_{\boldsymbol{\theta}\in K:|\boldsymbol{\theta}|_2\le 1} \boldsymbol{g}^T\boldsymbol{\theta}\right)^2, \tag{7.13}$$

cf. [1, Proposition 3.1]. The random variable (7.13) concentrates around its expectation. More precisely, [77, Lemma 4.9] combined with a Chernoff bound (as in [13, Lemma 20]) implies that with probability at least $1 - \alpha$, we have

$$|\Pi_K(\boldsymbol{g})|_2^2 \le \delta(K) + 2\sqrt{2\log(1/\alpha)\delta(K)} + 8\log(1/\alpha) \le 2\delta(K) + 10\log(1/\alpha). \tag{7.14}$$

We now define the intrinsic volumes of a polyhedral cone, which are closely related to the statistical dimension.

Definition 7.3 (Intrinsic volumes of a polyhedral cone)**.** Let $K \subset \mathbb{R}^n$ be a polyhedral cone and let $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, I_{n\times n})$. The intrinsic volumes of K are the real numbers

$$\nu_k(K) = \mathbb{P}\left(\Pi_K(\boldsymbol{g}) \text{ lies in the relative interior of a } k\text{-dimensional face of } K\right),$$

for all $k = 0, ..., n$.

The intrinsic volumes of a polyhedral cone K define a probability distribution on the discrete set $\{0, ..., n\}$. More precisely, define the random variable

$$V_K = \sum_{k=0}^{n} k\, \mathbf{1}_{\{\Pi_K(\boldsymbol{g}) \text{ lies in the relative interior of a } k\text{-dimensional face of } K\}}, \tag{7.15}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. The random variable V_K is valued in $\{0, ..., n\}$ and satisfies $\mathbb{P}(V_K = k) = \nu_k(K)$ for all $k = 0, ..., n$. The following identity was derived in [1, 77]:

$$\delta(K) = \sum_{k=0}^{n} k\nu_k(K), \tag{7.16}$$

that is, the statistical dimension $\delta(K)$ is the expectation of the random variable V_K. Furthermore, the random variable V_K concentrates around its expected value. The following concentration inequality is given in [77, Corollary 4.10]

$$\mathbb{P}\left(V_K - \delta(K) \geq \lambda\right) \leq \exp\left(-\frac{\delta(K)}{2} h\left(\frac{\lambda}{\delta(K)}\right)\right), \qquad \text{for all } \lambda > 0,$$

where $h(t) = (1+t)\log(1+t) - t$. Using the estimate $h^{-1}(t) \leq \sqrt{2t} + 3t$ (cf. [20, Corollary 12.12]), we obtain

$$\mathbb{P}\left(V_K - \delta(K) \geq 2\sqrt{x\delta(K)} + 6x\right) \leq \exp(-x), \qquad \text{for all } x > 0. \tag{7.17}$$

Deriving upper and lower bounds on the statistical dimension of a cone K may be a challenging problem. Some recipes to derive such bounds are proposed in [26, 1]. An exact formula is available for the statistical dimension of the cone $\mathcal{S}_n^\uparrow$ [1, (D.12)]. It is given by

$$\delta(\mathcal{S}_n^\uparrow) = \delta(\mathcal{S}_n^\downarrow) = \sum_{k=1}^{n} \frac{1}{k}, \tag{7.18}$$

so that $\log n \leq \delta(\mathcal{S}_n^\uparrow) \leq \log(en)$.

Finally, we will need the following characterization of the faces of the cone $\mathcal{S}_n^\uparrow$. The following proposition may be derived easily from the fact that if K is the polyhedron (7.11), and a face of K has the form (7.12).

Proposition 7.1. *Let $d \in \{1, ..., n\}$. The faces of dimension d of the cone $\mathcal{S}_n^\uparrow$ are the sets*

$$F(S) := \{\boldsymbol{u} = (u_1, ..., u_n)^T \in \mathcal{S}_n^\uparrow : u_{i-1} = u_i \text{ if } i \in S\}$$

where $S \subseteq \{2, ..., n\}$ with $|S| = n - d$. The cone $\mathcal{S}_n^\uparrow$ has no face of dimension 0.

Thus, for all $k = 1, ..., n$, the set $\mathcal{S}_n^\uparrow(k)$ is the union of all faces of dimension k.

7.3.3 Notation

For any $\boldsymbol{v} = (v_1, ..., v_n) \in \mathbb{R}^n$ and any $T \subset \{1, ..., n\}$, define the vector $\boldsymbol{v}_T \in \mathbb{R}^{|T|}$ as the restriction of $\boldsymbol{v}$ to T, that is,

$$\boldsymbol{v}_T = (v_{t_1}, ..., v_{t_{|T|}})^T \in \mathbb{R}^{|T|}$$

if $T = \{t_1, ..., t_{|T|}\}$ and $t_1 < ... < t_{|T|}$.

7.4 Adaptive confidence sets for nondecreasing sequences

The estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)$ projects $\mathbf{y}$ onto $\mathcal{S}_n^\uparrow$, so the vector $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)$ is nondecreasing. Let $\hat{k} = k(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)$ be the number of constant pieces of the Least Squares estimator. Using this notation, we define the statistic

$$\hat{r}_\uparrow = \frac{\sigma^2 \hat{k}(2 + 22\log(n) + 10\log(1/\alpha))}{n}. \tag{7.19}$$

Theorem 7.2. *For all $\alpha \in (0,1)$ and all $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$, the statistic $\hat{r}_{\uparrow}$ defined in (7.19) satisfies*

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \hat{r}_{\uparrow} \tag{7.20}$$

with probability at least $1-\alpha$.

The above proposition shows that the confidence set (7.7) with $\hat{r} = \hat{r}_{\uparrow}$ satisfies condition (7.3). Up to constants and logarithmic factors, the number of constant pieces $\hat{k}$ of the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ bounds the loss $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2$ from above with high probability. Since $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ can be computed in linear time, the integer $\hat{k}$ and the statistic $\hat{r}_{\uparrow}$ can also be computed in linear time. It is easy to compute $\hat{k}$ visually by drawing the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ and counting the number of jumps. The proof of Theorem 7.2 relies on concentration properties of the random variable (7.13).

Proof of Theorem 7.2. Let $s \leq t$ be two integers in $\{1,...,n\}$. Let

$$T_{s,t} = \{i = 1,...,n : s \leq i \leq t\} = \{s, s+1, ..., t-1, t\}, \tag{7.21}$$

that is, $T_{s,t}$ contains all consecutive integers from s to t. For any set T of the form (7.21), using the concentration property (7.14) of the random variable (7.13) with $K = \mathcal{S}_{|T|}^{\downarrow}$, we have with probability greater than $1-\alpha$,

$$|\Pi_{\mathcal{S}_{|T|}^{\downarrow}}(\boldsymbol{\xi}_T)|_2^2 \leq 2\delta\left(S_{|T|}^{\downarrow}\right) + 10\log(1/\alpha) \leq 2\log(en) + 10\log(1/\alpha),$$

where we used (7.18) for the last inequality. There are less that n^2 sets $T \subset \{1,...,n\}$ of the form (7.21). Using the union bound for all sets T of the form (7.21), we have $\mathbb{P}(\Omega(\alpha)) \geq 1-\alpha$ where

$$\Omega(\alpha) := \left\{ \sup_{T \in \{T_{s,t}, 1 \leq s \leq t \leq n\}} |\Pi_{\mathcal{S}_{|T|}^{\downarrow}}(\boldsymbol{\xi}_T)|_2^2 \leq \sigma^2 \left(2\log(en) + 10\log\left(\frac{n^2}{\alpha}\right)\right)\right\}.$$

Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ for notational simplicity. Then (7.6) with $\boldsymbol{u}$ replaced by $\boldsymbol{\mu}$ can be rewritten as

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \leq 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2.$$

By definition of $k(\cdot)$, there exists a partition $(\hat{T}_1, ..., \hat{T}_{\hat{k}})$ of $\{1,...,n\}$ such that $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ is constant on each $\hat{T}_j$, $j = 1,...,\hat{k}$. Furthermore, each $\hat{T}_j$ has the form (7.21). We have

$$\begin{aligned} 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 &= \sum_{j=1}^{\hat{k}} \left[2\boldsymbol{\xi}_{\hat{T}_j}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{\hat{T}_j} - |(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{T_j}|_2^2\right], \\ &\leq \sum_{j=1}^{\hat{k}} \left(\frac{\boldsymbol{\xi}_{\hat{T}_j}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{\hat{T}_j}}{|(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{\hat{T}_j}|_2}\right)^2, \end{aligned}$$

where have used the elementary inequality $2ab - a^2 \leq b^2$. By definition of $(\hat{T}_1, ..., \hat{T}_{\hat{k}})$, $\hat{\boldsymbol{\mu}}$ is constant on $\hat{T}_j$ for each $j = 1,...,\hat{k}$, thus the subsequence is non-increasing: $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})_{\hat{T}_j} \in \mathcal{S}_{|\hat{T}_j|}^{\downarrow}$. By taking the supremum, we obtain

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \leq \sum_{j=1}^{\hat{k}} \sup_{\boldsymbol{v} \in \mathcal{S}_{|\hat{T}_j|}^{\downarrow} : |\boldsymbol{v}|_2^2 \leq 1} (\boldsymbol{\xi}_{\hat{T}_j}^T \boldsymbol{v})^2 = \sum_{j=1}^{\hat{k}} |\Pi_{\mathcal{S}_{|\hat{T}_j|}^{\downarrow}}(\boldsymbol{\xi}_{|\hat{T}_j|})|_2^2,$$

where we used (7.13) for the last equality. On the event $\Omega(\alpha)$ and by definition of $\hat{r}_\uparrow$,

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \le \sum_{j=1}^{\hat{k}} |\Pi_{\mathcal{S}^\downarrow_{|\hat{T}_j|}}(\boldsymbol{\xi}_{|\hat{T}_j|})|_2^2 \le \sigma^2 \hat{k} \left(2\log(en) + 10\log(n^2/\alpha)\right) = n\hat{r}_\uparrow.$$

□

We have established the existence of an honest confidence interval of the form

$$\hat{C}_n := \{\boldsymbol{v} \in \mathcal{S}_n^\uparrow : \quad \|\boldsymbol{v} - \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow)\|^2 \le \hat{r}\}.$$

This confidence set has uniform coverage over all $\boldsymbol{\mu} \in \mathcal{S}_n^\uparrow$, i.e., it satisfies (7.3). The next result implies that the diameter of this confidence set is minimax optimal up to logarithmic factors.

Theorem 7.3. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^\uparrow$ and $\gamma \in (0,1)$. The random variable $\hat{k} = k(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^\uparrow))$ satisfies*

$$\hat{k} \le 2k(\boldsymbol{\mu}) \log\left(\frac{en}{k(\boldsymbol{\mu})}\right) + 7\log(1/\gamma)) \tag{7.22}$$

with probability greater than $1-\gamma$. Furthermore,

$$\mathbb{E}_{\boldsymbol{\mu}}[\hat{k}] \le k(\boldsymbol{\mu}) \log\left(\frac{en}{k(\boldsymbol{\mu})}\right).$$

Proof of Theorem 7.3. Let $k = k(\boldsymbol{\mu})$ and let $(T_1, ..., T_k)$ be a partition of $\{1, ..., n\}$ such that for all $j = 1, ..., k$, $\boldsymbol{\mu}_{T_j}$ is constant. As $\boldsymbol{\mu}$ is nondecreasing, T_j has the form (7.21) for all $j = 1, ..., k$. Define the closed convex cone

$$K = \mathcal{S}^\uparrow_{|T_1|} \times \mathcal{S}^\uparrow_{|T_2|} \times ... \times \mathcal{S}^\uparrow_{|T_k|} \subset \mathbb{R}^n$$

and let $\hat{\boldsymbol{\mu}}^* = \Pi_K(\mathbf{y})$. It is clear that

$$\min_{\boldsymbol{u}\in K} \sum_{j=1}^k |\mathbf{y}_{T_j} - \boldsymbol{u}_{T_j}|_2^2 = \min_{\boldsymbol{u}_1 \in \mathcal{S}^\uparrow_{|T_1|}, ..., \boldsymbol{u}_k \in \mathcal{S}^\uparrow_{|T_k|}} \sum_{j=1}^k |\mathbf{y}_{T_j} - \boldsymbol{u}_{T_j}|_2^2 = \sum_{j=1}^k \min_{\boldsymbol{u}_j \in \mathcal{S}^\uparrow_{|T_j|}} |\mathbf{y}_{T_j} - \boldsymbol{u}_{T_j}|_2^2.$$

Thus, as $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\xi}$ and $\boldsymbol{\mu}$ is constant on each T_j, we have

$$\hat{\boldsymbol{\mu}}^*_{T_j} = \Pi_{\mathcal{S}^\uparrow_{|T_j|}}(\mathbf{y}_{T_j}) = \Pi_{\mathcal{S}^\uparrow_{|T_j|}}(\boldsymbol{\mu}_{T_j} + \boldsymbol{\xi}_{T_j}) = \boldsymbol{\mu}_{T_j} + \Pi_{\mathcal{S}^\uparrow_{|T_j|}}(\boldsymbol{\xi}_{T_j}).$$

As adding the constant sequence $\boldsymbol{\mu}_{T_j}$ does not modify the number of constant pieces (or the number of jumps), we have

$$k\left(\hat{\boldsymbol{\mu}}^*_{T_j}\right) = k\left(\boldsymbol{\mu}_{T_j} + \Pi_{\mathcal{S}^\uparrow_{|T_j|}}(\boldsymbol{\xi}_{T_j})\right) = k\left(\Pi_{\mathcal{S}^\uparrow_{|T_j|}}(\boldsymbol{\xi}_{T_j})\right) = k\left((\Pi_K(\boldsymbol{\xi}))_{T_j}\right).$$

Let V_K be the random variable defined in (7.15). By the properties of product cones given in [77, Section 5.2], V_K has the same distribution as

$$V_{\mathcal{S}^\uparrow_{|T_1|}} + ... + V_{\mathcal{S}^\uparrow_{|T_k|}}.$$

By Proposition 7.1, for all $j = 1, ..., k$ we have $k(\hat{\boldsymbol{\mu}}^*_{T_j}) = V_{\mathcal{S}^\uparrow_{|T_j|}}$ so that $\sum_{j=1}^k k(\hat{\boldsymbol{\mu}}^*_{T_j})$ is distributed as V_K. By (7.16), $\mathbb{E}V_K = \delta(K)$ and by (7.17), with probability greater than $1 - \gamma$ we have

$$V_K \leq 2\delta(K) + 7\log(1/\gamma).$$

To bound $\delta(K)$ from above, we use that the statistical dimension of a direct product of cones is the sum of the statistical dimensions (cf. [1, Proposition 3.1])

$$\delta(K) = \sum_{j=1}^k \delta(\mathcal{S}^\uparrow_{|T_j|}) \leq \sum_{j=1}^k \log(e|T_j|) \leq k\log(en/k),$$

where we have used (7.18) and Jensen's inequality.

The random variable V_K is distributed as $\sum_{j=1}^k k(\hat{\boldsymbol{\mu}}^*_{T_j})$. Thus, to complete the proof, it is enough to prove that almost surely, $\hat{k} := k(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^\uparrow_n)) \leq \sum_{j=1}^k k(\hat{\boldsymbol{\mu}}^*_{T_j})$. Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^\uparrow_n)$ for notational simplicity. It is clear that

$$k(\hat{\boldsymbol{\mu}}) = |\{\hat{\mu}_i, i = 1, ..., n\}| \leq \sum_{j=1}^k k(\hat{\boldsymbol{\mu}}_{T_j}) = \sum_{j=1}^k |\{\hat{\mu}_i, i \in T_j\}|,$$

since a piece counted on the left hand must be counted at least once on the right hand side. For all $j = 1, ..., k$, $\hat{\boldsymbol{\mu}}_{T_j}$ and $\hat{\boldsymbol{\mu}}^*_{T_j}$ are solutions of the minimization problems

$$\hat{\boldsymbol{\mu}}^*_{T_j} = \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{S}^\uparrow_{|T_j|}} |\boldsymbol{v} - \mathbf{y}_{T_j}|_2^2, \qquad \hat{\boldsymbol{\mu}}_{T_j} = \operatorname*{argmin}_{\substack{\boldsymbol{v} \in \mathcal{S}^\uparrow_{|T_j|}: \\ \hat{\boldsymbol{\mu}}_{\min(T_j)-1} \leq \boldsymbol{v}_1, \\ \boldsymbol{v}_{|T_j|} \leq \hat{\boldsymbol{\mu}}_{\max(T_j)+1}}} |\boldsymbol{v} - \mathbf{y}_{T_j}|_2^2,$$

where by convention $\hat{\boldsymbol{\mu}}_0 = -\infty$ and $\hat{\boldsymbol{\mu}}_{n+1} = +\infty$. This means that $\hat{\boldsymbol{\mu}}_{T_j}$ is solution of a minimization problem with additional constraints at the boundary. By Lemma 7.13, we have

$$k\left(\hat{\boldsymbol{\mu}}_{T_j}\right) \leq k\left(\hat{\boldsymbol{\mu}}^*_{T_j}\right)$$

for all $j = 1, ..., k$, which completes the proof. □

Corollary 7.4. *Let $J = \{1, ..., n\}$ and define the collection of models $(E_k)_{k \in J} = \left(\mathcal{S}^\uparrow_n(k)\right)_{k \in J}$. The random variable $\hat{r}_\uparrow$ defined in (7.19) satisfies (7.20), (7.8) and (7.9) with $\hat{r}$ replaced by $\hat{r}_\uparrow$. Thus, the ball centered at $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^\uparrow_n)$ of radius $\sqrt{\hat{r}_\uparrow}$ is an honest confidence set, which is adaptive in probability and in expectation with respect to the models $\left(\mathcal{S}^\uparrow_n(k)\right)_{k=1,...,n}$.*

7.5 Nondecreasing sequences with bounded total variation

Let $V > 0$. If the unknown parameter $\boldsymbol{\mu}$ satisfies $\mu_n - \mu_1 \leq V$, the risk of the Least Squares estimator satisfy [107, (28)]

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}^\uparrow_n) - \boldsymbol{\mu}\|^2 \leq \sigma^2\kappa^2 \left(\left(\frac{V}{\sigma n} \right)^{2/3} + \frac{\log(en)}{n} \right),$$

where $\kappa \leq 3.6$. Thus, an explicit constant is readily available [107, (2.8)]. It is possible to deduce from this risk bound an upper bound on the loss $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2$ with high probability. We proceed as follows.

The function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by $f(\boldsymbol{v}) = \|\Pi_{\mathcal{S}_n^{\uparrow}}(\boldsymbol{\mu} + \sigma\boldsymbol{v}) - \boldsymbol{\mu}\|$ is Lipschitz with coefficient $\sigma/\sqrt{n}$ as for all $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^n$,

$$|f(\boldsymbol{v}) - f(\boldsymbol{v}')| \leq \|\Pi_{\mathcal{S}_n^{\uparrow}}(\boldsymbol{\mu} + \sigma\boldsymbol{v}) - \Pi_{\mathcal{S}_n^{\uparrow}}(\boldsymbol{\mu} + \sigma\boldsymbol{v}')\| \leq \sigma\|\boldsymbol{v} - \boldsymbol{v}'\| = (\sigma/\sqrt{n})|\boldsymbol{v} - \boldsymbol{v}'|_2. \tag{7.23}$$

By the Gaussian concentration inequality [20, Theorem 5.6], the following holds with probability greater than $1 - \alpha$

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\| \leq \mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\| + \sigma\sqrt{\frac{2\log(1/\alpha)}{n}}.$$

Using that $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain the following for all $\alpha \in (0,1)$: If $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ and $\mu_n - \mu_1 \leq V$, then

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq 2\kappa^2\sigma^2\left(\frac{V}{\sigma n}\right)^{2/3} + \frac{2\kappa^2\sigma^2\log(en) + 4\sigma^2\log(1/\alpha)}{n}$$

with probability greater than $1 - \alpha$.

Let $V_{\boldsymbol{\mu}} = \mu_n - \mu_1$ and $\hat{V} = y_n - y_1$. The random variable $\hat{V} - V_{\boldsymbol{\mu}}$ is centered Gaussian with variance $2\sigma^2$, so that

$$V_{\boldsymbol{\mu}} \leq \hat{V} + 2\sigma\sqrt{\log(1/\alpha)}$$

with probability greater than $1 - \alpha$. Thus, we have established the following.

Proposition 7.5. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$. Define the statistic $\hat{s}_{\uparrow}$ by*

$$\sqrt{\hat{s}_{\uparrow}} = 2\kappa^2\sigma^2\left(\frac{\hat{V} + 2\sigma\sqrt{\log(1/\alpha)}}{\sigma n}\right)^{2/3} + \frac{2\kappa^2\sigma^2\log(en) + 4\sigma^2\log(1/\alpha)}{n}$$

where $\kappa \leq 3.6$ is the constant from [107] that appears in (7.23). Then we have $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \hat{s}_{\uparrow}$ with probability greater than $1 - \alpha$.

Furthermore, it is clear that $\hat{V} \leq V_{\boldsymbol{\mu}} + 2\sigma\sqrt{\log(1/\gamma)}$ with probability greater than $1 - \gamma$ for all $\gamma \in (0,1)$.

Proposition 7.6. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ and let $V = \mu_n - \mu_1$. Then the statistic $\hat{s}_{\uparrow}$ defined above satisfies*

$$\hat{s}_{\uparrow} \leq 2\kappa^2\sigma^2\left(\frac{\hat{V} + 2\sigma\sqrt{\log(1/(\gamma\alpha))}}{\sigma n}\right)^{2/3} + \frac{2\kappa^2\sigma^2\log(en) + 4\sigma^2\log(1/\alpha)}{n} \tag{7.24}$$

with probability at least $1 - \gamma$ for all $\gamma \in (0,1)$.

Theorem 7.7. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$. The statistic $\min(\hat{r}_{\uparrow}, \hat{s}_{\uparrow})$ satisfies*

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \min(\hat{r}_{\uparrow}, \hat{s}_{\uparrow})$$

with probability at least $1 - 2\alpha$. Furthermore, for all $\gamma \in (0,1)$, the statistic $\min(\hat{r}_{\uparrow}, \hat{s}_{\uparrow})$ is bounded from above with probability at least $1 - 2\gamma$, by the minimum of the right hand side of (7.22) and the right hand side of (7.24).

For all $V \geq \sigma$ and all $k = 1, ..., n$, define the class

$$\mathcal{S}_n^{\uparrow}(k, V) := \{\boldsymbol{v} = (v_1, ..., v_n)^T \in \mathcal{S}_n^{\uparrow} : \quad k(\boldsymbol{v}) \leq k \text{ and } v_n - v_1 \leq V\}.$$

For small enough $\alpha > 0$, the quantity $R_\alpha^*(\mathcal{S}_n^{\uparrow}(k, V))$ defined in (7.2) is greater than

$$c\sigma^2 \min\left(\left(\frac{V}{\sigma n}\right)^{2/3}, \frac{k}{n}\right)$$

for some absolute constant $c > 0$, cf. [13, Proposition 4]. Thus, the statistic $\min(\hat{r}_{\uparrow}, \hat{s}_{\uparrow})$ of Theorem 7.7 induces an honest confidence ball centered at the Least Squares estimator, and this confidence ball is adaptive in probability for the collection of models

$$(\mathcal{S}_n^{\uparrow}(k, V))_{k \in \{1,...,n\}, V \in [\sigma, +\infty)}.$$

7.6 Adaptive confidence sets for convex sequences

Confidence sets can also be obtained in univariate convex regression. If $n \geq 3$, define the set of convex sequences $\mathcal{S}_n^{\cup}$ by

$$\mathcal{S}_n^{\cup} := \{\boldsymbol{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n : \ 2u_i \leq u_{i+1} + u_{i-1}, \ i = 2, \ldots, n-1\},$$

and define $\mathcal{S}_n^{\cup} = \mathbb{R}$ if $n = 1$ and $\mathcal{S}_n^{\cup} = \mathbb{R}^2$ if $n = 2$. For all $n \geq 1$, define the cone of concave sequences by $\mathcal{S}_n^{\cap} := -\mathcal{S}_n^{\cup}$.

For any $\boldsymbol{u} \in \mathcal{S}_n^{\cup}$, let $q(\boldsymbol{u}) - 1 \geq 0$ be the number of inequalities $2u_i \leq u_{i+1} + u_{i-1}, i = 2, ..., n-1$ that are strict. The integer $q(\boldsymbol{u})$ is also the smallest positive integer such that $\boldsymbol{u}$ is piecewise affine with $q(\boldsymbol{u})$ pieces. Define the sets

$$\mathcal{S}_n^{\cup}(q) = \{\boldsymbol{u} \in \mathcal{S}_n^{\cup} : q(\boldsymbol{u}) \leq q\}, \qquad q = 1, ..., n-1.$$

The set $\mathcal{S}_n^{\cup}(1)$ is the subspace of all affine sequences while $\mathcal{S}_n^{\cup}(2), ..., \mathcal{S}_n^{\cup}(n-2)$ are closed non-convex sets. We have

$$\mathcal{S}_n^{\cup}(1) \subset \mathcal{S}_n^{\cup}(2) \subset ... \subset \mathcal{S}_n^{\cup}(n-1) = \mathcal{S}_n^{\cup}.$$

These sets represent models of increasing complexity.

There exist numerical constants $c, c' > 0$ such that for all $\alpha \leq (0, \min(c, 1))$ and any $q = 1, ..., n-1$, we have

$$\frac{c'\sigma^2 q}{n} \leq R_\alpha^*(\mathcal{S}_n^{\cup}(q)) \leq \frac{20\sigma^2 q \log(en/q)}{n} + \frac{10\log(1/\alpha)}{n}, \tag{7.25}$$

cf. [13, Theorem 6] for the upper bound and [14, Proposition 7] for the lower bound. Thus, for $\alpha > 0$ small enough, the quantity $R_\alpha^*(\mathcal{S}_n^{\cup}(q))$ is of order $q\sigma^2/n$, up to logarithmic factors.

The statistical dimension of the cone $\mathcal{S}_n^{\cup}$ satisfies [13]

$$\delta(\mathcal{S}_n^{\cup}) = \delta(\mathcal{S}_n^{\cap}) \leq 10\log(en). \tag{7.26}$$

It is not known whether this upper bound is sharp. However, the fact that the statistical dimension of $\mathcal{S}_n^{\cup}$ grows slower that a logarithmic function of n is enough for the purpose of the present paper.

The following bound on the risk of $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\cup})$ will be useful.

Proposition 7.8 ([13]). *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$. Then*

$$\mathbb{E}_{\boldsymbol{\mu}} |\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}) - \boldsymbol{\mu}|_2^2 \leq \mathbb{E}_{\boldsymbol{\mu}} \left[\Big(\sup_{\boldsymbol{v} \in \mathcal{T}_{\boldsymbol{\mu}} : |\boldsymbol{v}|_2 \leq 1} \boldsymbol{\xi}^T \boldsymbol{v} \Big)^2 \right] = \sigma^2 \delta(\mathcal{T}_{\boldsymbol{\mu}}) \leq 10 \sigma^2 q(\boldsymbol{\mu}) \log \frac{en}{q(\boldsymbol{\mu})},$$

where $\mathcal{T}_{\boldsymbol{\mu}}$ is the tangent cone at $\boldsymbol{\mu}$ defined by

$$\mathcal{T}_{\boldsymbol{\mu}} := \mathit{closure}\{t(\boldsymbol{u} - \boldsymbol{\mu}), t \geq 0, \boldsymbol{u} \in \mathcal{S}_n^{\cup}\}.$$

An outline of the proof of this result is as follows. More details may be found in [13].

Outline of the proof of Proposition 7.8. The inequality $\mathbb{E}_{\boldsymbol{\mu}} |\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}) - \boldsymbol{\mu}|_2^2 \leq \sigma^2 \delta(\mathcal{T}_{\boldsymbol{\mu}})$ was proved by [84], it is a direct consequence of (7.6) with $\boldsymbol{u} = \boldsymbol{\mu}$. To bound from above the statistical dimension of $\mathcal{T}_{\boldsymbol{\mu}}$, we have the inclusion

$$\mathcal{T}_{\boldsymbol{\mu}} \subset \mathcal{S}_{|T_1|}^{\cup} \times ... \times \mathcal{S}_{|T_{q(\boldsymbol{\mu})}|}^{\cup},$$

where $(T_1, ..., T_{q(\boldsymbol{\mu})})$ is a partition of $\{1, ..., n\}$ such that $\boldsymbol{\mu}$ is affine on each T_j, $j = 1, ..., q(\boldsymbol{\mu})$. The formula for the statistical dimension of a direct product of cones [1, Proposition 3.1] yields

$$\delta(\mathcal{S}_{|T_1|}^{\cup} \times ... \times \mathcal{S}_{|T_{q(\boldsymbol{\mu})}|}^{\cup}) = \sum_{j=1}^{q(\boldsymbol{\mu})} \delta(\mathcal{S}_{|T_j|}^{\cup}) \leq 10 \sum_{j=1}^{q(\boldsymbol{\mu})} \log(e|T_j|) \leq 10 \log(en/q(\boldsymbol{\mu})),$$

where we used (7.26) and Jensen's inequality. □

We now turn to the construction of confidence sets. Recall that if $\boldsymbol{u} \in \mathcal{S}_n^{\cup}$ is a convex sequence, $q(\boldsymbol{u})$ is the number of pieces in the piecewise affine decomposition of $\boldsymbol{u}$. Let $\hat{q} := q(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ be the number of affine pieces of the Least Squares estimator. Then, define the statistic

$$\hat{r}_{\cup} = \frac{\sigma^2 \hat{q}(20 + 40 \log(n) + 10 \log(1/\alpha))}{n}. \tag{7.27}$$

Similarly to the case of the statistic $\hat{r}_{\uparrow}$ in isotonic regression, the following result shows that the confidence ball (7.7) with $\hat{r} = \hat{r}_{\cup}$ enjoys uniform coverage over all $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$.

Theorem 7.9. *For all $\alpha \in (0, 1)$ and all $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$, the statistic $\hat{r}_{\cup}$ defined in (7.19) satisfies*

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}) - \boldsymbol{\mu}\|^2 \leq \hat{r}_{\cup}, \tag{7.28}$$

with probability at least $1 - \alpha$.

The above result is analog to Theorem 7.2. The numerical constants are slightly worse in the case of the present section because the upper bound (7.26) on the statistical dimension of the cone $\mathcal{S}_n^{\cup}$ is slightly worse than (7.18). The proof of Theorem 7.9 is similar to the proof of Theorem 7.2 and can be found in the appendix.

Now, the goal is to show that the statistic $\hat{r}_{\cup}$ is of the same order as the minimax quantity (7.25). We employ a different strategy than in the previous section.

For any function $g : \mathbb{R}^n \to \mathbb{R}^n$ which is weakly differentiable, the divergence of g is the random variable

$$D_g(\mathbf{y}) = \sigma^2 \sum_{i=1}^{n} \frac{\partial}{\partial y_i} g(\mathbf{y})_i.$$

It is well known that by Stein's identity, under suitable conditions on g (cf. [80, Section 2] or [99, Lemma 3.6]), we have

$$\sigma^2 \mathbb{E}_{\boldsymbol{\mu}} D_g(\mathbf{y}) = \mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{\xi}^T g(\mathbf{y})]. \tag{7.29}$$

The divergence of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}) = \Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y})$ is given in [30, Proposition 2.7] (see also [80]). Namely, we have the following result.

Proposition 7.10 ([80, 30]). *If $g(\cdot) = \Pi_{\mathcal{S}_n^{\cup}}(\cdot)$ is the projection onto the cone of convex sequences, then (7.29) holds and we have*

$$D_g(\mathbf{y}) = \hat{q} + 1$$

almost surely, where $\hat{q} = q(\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup}))$.

This result can be used to bound from above the expected radius of the statistic $\hat{r}_{\cup}$.

Theorem 7.11. *Let $\boldsymbol{\mu} \in \mathcal{S}_n^{\cup}$. Then*

$$\mathbb{E}_{\boldsymbol{\mu}}[\hat{q}] \leq 10 q(\boldsymbol{\mu}) \log \frac{en}{q(\boldsymbol{\mu})} - 1. \tag{7.30}$$

Furthermore, for all $\alpha \in (0,1)$, the statistic (7.27) satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}[\hat{r}_{\cup}] \leq \frac{\sigma^2 q(\boldsymbol{\mu}) \operatorname{polylog}(n, 1/\alpha)}{n}. \tag{7.31}$$

where $\operatorname{polylog}(n, 1/\alpha) = 10 \log(en)(20 + 40 \log(n) + 10 \log(1/\alpha))$.

Proof. By Proposition 7.10 and (7.29), we have

$$\sigma^2 \mathbb{E}_{\boldsymbol{\mu}}[1 + \hat{q}] = \mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{\xi}^T \Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y})] = \mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{\xi}^T (\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu})].$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\sigma^2 \mathbb{E}_{\boldsymbol{\mu}}[1 + \hat{q}] &\leq \mathbb{E}_{\boldsymbol{\mu}}^{1/2} \left[\left(\frac{\boldsymbol{\xi}^T (\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu})}{|\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu}|_2} \right)^2 \right] \mathbb{E}_{\boldsymbol{\mu}}^{1/2} |\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu}|_2^2, \\
&\leq \sigma \sqrt{\delta(\mathcal{T}_{\boldsymbol{\mu}})} \mathbb{E}_{\boldsymbol{\mu}}^{1/2} |\Pi_{\mathcal{S}_n^{\cup}}(\mathbf{y}) - \boldsymbol{\mu}|_2^2.
\end{aligned}$$

Using Proposition 7.8 completes the proof of (7.30). Inequality (7.31) is a direct consequence of (7.30) and of the definition of $\hat{r}_{\cup}$. □

The above result is different from Theorem 7.3 in isotonic regression. Theorem 7.3 controls both the expectation and the deviations of $\hat{k}$. In this section, Theorem 7.11 only controls the expectation of $\hat{q}$. This comes from the use of Stein's identity in the proof of Theorem 7.11, which yields a result only in expectation.

The arguments used to prove Theorem 7.3 are based on the concentration properties of the intrinsic volumes of cones, while the proof of Theorem 7.11 relies on Stein's identity and Proposition 7.10. Thus, we have presented two methods to bound from above the expected diameter of the confidence sets constructed in the present paper.

Corollary 7.12. *Let $J = \{1, ..., n-1\}$ and define the collection of models $(E_k)_{k \in J} = (\mathcal{S}_n^{\cup}(k))_{k=1,...,n-1}$. The random variable $\hat{r}_{\cup}$ defined in (7.27) satisfies (7.28) and (7.9) with $\hat{r}$ replaced by $\hat{r}_{\uparrow}$. Thus, the ball centered at $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ of radius $\sqrt{\hat{r}_{\uparrow}}$ is an honest confidence set, which is adaptive in expectation with respect to the models $(\mathcal{S}_n^{\cup}(k))_{k=1,...,n-1}$.*

7.7 Concluding remarks

We have shown that it is possible to design honest and adaptive confidence sets for the estimation problem over two convex polyhedra: the cone of nondecreasing sequences and the cone of non-increasing sequences. The confidence sets defined in the previous sections adapt automatically to the unknown dimension of the smallest face that contains the true parameter $\boldsymbol{\mu}$. Theorems 7.2, 7.3, 7.9 and 7.11 provide a deeper understanding of the statistical complexity of these polyhedra in the case where the true parameter $\boldsymbol{\mu}$ lies on a low-dimensional face.

Let K be either $\mathcal{S}_n^{\uparrow}$ or $\mathcal{S}_n^{\cup}$ and let us summarize some statistical properties of the Least Squares estimator around low-dimension faces.

1. If the true parameter $\boldsymbol{\mu}$ belongs to a d-dimensional face of K, then the rate of convergence of the Least Squares estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(K)$ is almost parametric, of order $\sigma^2 d/n$ [29, 50], and it is the minimax rate up to logarithmic factors.

2. If the true parameter $\boldsymbol{\mu}$ is well approximated by some $\boldsymbol{u} \in K$ and $\boldsymbol{u}$ lies in a d-dimensional face, then the rate of the estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(K)$ is still parametric of order $\sigma^2 d/n$ up to logarithmic factors. This phenomenon takes the form of oracle inequalities [29, 50, 13]. Furthermore, these bounds hold both in expectation and with high probability [13].

3. Let $\alpha \in (0,1)$. By Theorems 7.2, 7.3, 7.9 and 7.11, there exists a $(1-\alpha)$ confidence set $\hat{C}_n$ which depends only on K, σ and α such that the following holds. For all $d = 1, ..., n$ and for all $\boldsymbol{\mu} \in K$, if the true parameter $\boldsymbol{\mu}$ belongs to a d-dimensional face of K, then the diameter of $\hat{C}_n$ is of order $\sigma^2 d/n$ up to logarithmic factors.

These results illustrate that an remarkable statistical phenomenon appears for the estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(K)$ if the true parameter lies around a low-dimensional face of K: In that case the estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(K)$ converges at an almost parametric rate, and it is possible to construct confidence sets whose radius is of the same order as this almost parametric rate.

A natural question is whether these results can be extended to other polyhedra. Are there other examples polyhedra K for which this phenomenon appears? Is it possible to generalize these results to a large class of polyhedra? To our knowledge, there is no general method to construct adaptive confidence sets such as the ones studied in Sections 7.4 to 7.6 of the present paper. A generalization of (1) and (2) is the following oracle inequality. For any closed convex set K and any $\boldsymbol{\mu} \in \mathbb{R}^n$, we have

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in K} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 \delta\left(\mathcal{C}_{\boldsymbol{u},K}\right)}{n} \right)$$

where $\mathcal{C}_{\boldsymbol{u},K}$ is the tangent cone at $\boldsymbol{u}$ defined by $\mathcal{C}_{\boldsymbol{u},K} = \{\boldsymbol{v} - t\boldsymbol{u}, \boldsymbol{v} \in K, t \geq 0\}$ (cf. [84] in the well-specified case and [13] in the miss-specified). A similar oracle inequality holds with high probability using the concentration inequality (7.14) from [1]. Namely, for all $x > 0$ we have [13]

$$\|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in K} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2}{n} \left(\delta\left(\mathcal{C}_{\boldsymbol{u},K}\right) + 2\sqrt{2x\delta\left(\mathcal{C}_{\boldsymbol{u},K}\right)} + 8x \right) \right)$$

with probability at least $1 - e^{-x}$. In the well-specified case, taking $\boldsymbol{u} = \boldsymbol{\mu}$ in (7.32) we obtain

$$\mathbb{E}\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \boldsymbol{\mu}\|^2 \leq \frac{\sigma^2 \delta\left(\mathcal{C}_{\boldsymbol{\mu},K}\right)}{n}.$$

It was proved in [84] that this risk bound becomes tight as the noise level σ tends to 0. If K is a polyhedron and if $\boldsymbol{\mu}, \boldsymbol{\mu}'$ belong to the relative interior of the same face F of K, then the tangent cones are the same, that is, $\mathcal{C}_{\boldsymbol{\mu},K} = \mathcal{C}_{\boldsymbol{\mu}',K}$. This suggests that the statistical dimension of the tangent cone $\delta(\mathcal{C}_{\boldsymbol{\mu},K})$ is an insightful statistical invariant of the face F.

7.8 Appendix: Technical Lemma

Lemma 7.13. *In the present Lemma, all quantities are deterministic. Let $a \in [-\infty, +\infty)$ and $b \in (-\infty, +\infty]$ such that $a \leq b$. Let $\mathbf{y} \in \mathbb{R}^n$. Define $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ as the unique solutions of the minimization problems*

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{S}_n^{\uparrow}} |\mathbf{y} - \boldsymbol{v}|_2^2, \tag{7.32}$$

$$\boldsymbol{\theta} \in \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{S}_n^{\uparrow}(a,b)} |\mathbf{y} - \boldsymbol{v}|_2^2 \tag{7.33}$$

where $\mathcal{S}_n^{\uparrow}(a,b) \coloneqq \{\boldsymbol{v} = (v_1, ..., v_n)^T \in \mathcal{S}_n^{\uparrow} : a \leq v_1, v_n \leq b\}$. Then $k(\boldsymbol{\theta}) \leq k(\boldsymbol{\theta}^*)$.

The intuition behind this Lemma is that if a constraint is not saturated for $\boldsymbol{\theta}$, this constraint is not saturated for $\boldsymbol{\theta}^*$ either, so $\boldsymbol{\theta}^*$ has at least as many jumps as $\boldsymbol{\theta}$.

Proof of Lemma 7.13. Let $T_a = \{i = 1, ..., n : \hat{\theta}_i^* \leq a\}$, $T_b = \{i = 1, ..., n : \hat{\theta}_i^* \geq b\}$ and $T_c = \{i = 1, ..., n : a < \hat{\theta}_i^* < b\}$. We will prove that the unique minimizer $\boldsymbol{\theta}$ of the problem (7.33) is

$$\boldsymbol{\theta}_{T_a} = a\mathbf{1}_{T_a}, \qquad \boldsymbol{\theta}_{T_c} = \boldsymbol{\theta}^*_{T_c}, \qquad \boldsymbol{\theta}_{T_b} = b\mathbf{1}_{T_b}, \tag{7.34}$$

where $\mathbf{1} = (1, ..., 1)^T \in \mathbb{R}^n$. Then it is clear that $k(\boldsymbol{\theta}) = 1 + k(\boldsymbol{\theta}^*_{T_c}) + 1 \leq k(\boldsymbol{\theta}^*_{T_a}) + k(\boldsymbol{\theta}^*_{T_c}) + k(\boldsymbol{\theta}^*_{T_b}) = k(\boldsymbol{\theta}^*)$.

First, by strong convexity there exists a unique solution to the minimization problem (7.33), and a unique solution to the minimization problem (7.32). Second, by the characterization of the projection onto the closed convex set $\mathcal{S}_n^{\uparrow}(a,b)$, if $\boldsymbol{\theta}$ satisfies

$$A_{\boldsymbol{u}} \coloneqq (\boldsymbol{u} - \boldsymbol{\theta})^T(\mathbf{y} - \boldsymbol{\theta}) \leq 0$$

for all $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}(a,b)$, then $\boldsymbol{\theta}$ is the unique solution to the minimization problem (7.33). Let $\boldsymbol{\theta}$ be defined by (7.34). By simple algebra, for all $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}(a,b)$,

$$\begin{aligned} A_{\boldsymbol{u}} =& (\boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a} + \boldsymbol{\theta}^*_{T_a} - \boldsymbol{\theta}^*_{T_a})^T(\mathbf{y}_{T_a} - \boldsymbol{\theta}^*_{T_a}) + (\boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a})^T(\boldsymbol{\theta}^*_{T_a} - a\mathbf{1}_{T_a}) \\ &+ (\boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b} + \boldsymbol{\theta}^*_{T_b} - \boldsymbol{\theta}^*_{T_b})^T(\mathbf{y}_{T_b} - \boldsymbol{\theta}^*_{T_b}) + (\boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b})^T(\boldsymbol{\theta}^*_{T_b} - b\mathbf{1}_{T_b}) \\ &+ (\boldsymbol{u}_{T_c} - \boldsymbol{\theta}^*_{T_c})^T(\mathbf{y}_{T_c} - \boldsymbol{\theta}^*_{T_c}). \end{aligned}$$

If a vector $\boldsymbol{v}$ has nonnegative entries and a vector $\boldsymbol{x}$ have non-positive entries, then $\boldsymbol{v}^T\boldsymbol{x} \leq 0$, so $(\boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a})^T(\boldsymbol{\theta}^*_{T_a} - a\mathbf{1}_{T_a}) \leq 0$ and $(\boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b})^T(\boldsymbol{\theta}^*_{T_b} - b\mathbf{1}_{T_b}) \leq 0$. Thus,

$$\begin{aligned} A_{\boldsymbol{u}} \leq& (\boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a} + \boldsymbol{\theta}^*_{T_a} - \boldsymbol{\theta}^*_{T_a})^T(\mathbf{y}_{T_a} - \boldsymbol{\theta}^*_{T_a}) \\ &+ (\boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b} + \boldsymbol{\theta}^*_{T_b} - \boldsymbol{\theta}^*_{T_b})^T(\mathbf{y}_{T_b} - \boldsymbol{\theta}^*_{T_b}) \\ &+ (\boldsymbol{u}_{T_c} - \boldsymbol{\theta}^*_{T_c})^T(\mathbf{y}_{T_c} - \boldsymbol{\theta}^*_{T_c}), \end{aligned}$$

and the right hand side of the previous display is equal to

$$(\boldsymbol{v}-\boldsymbol{\theta}^*)^T(\mathbf{y}-\boldsymbol{\theta}^*), \tag{7.35}$$

where $\boldsymbol{v}\in\mathbb{R}^n$ is defined by

$$\boldsymbol{v}_{T_a} \coloneqq \boldsymbol{u}_{T_a} - a\mathbf{1}_{T_a} + \boldsymbol{\theta}^*_{T_a}, \qquad \boldsymbol{v}_{T_c} \coloneqq \boldsymbol{u}_{T_c}, \qquad \boldsymbol{v}_{T_b} \coloneqq \boldsymbol{u}_{T_b} - b\mathbf{1}_{T_b} + \boldsymbol{\theta}^*_{T_b}.$$

We have $\boldsymbol{v}\in\mathcal{S}_n^{\uparrow}$ by definition of T_a, T_c and T_b. The quantity (7.35) is non-positive because $\boldsymbol{\theta}^*$ is the projection of $\mathbf{y}$ onto the convex set $\mathcal{S}_n^{\uparrow}$. Thus we have established that $A_{\boldsymbol{u}}\le 0$ for all $\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}(a,b)$, so that the unique solution of the minimization problem (7.33) is given by the expression (7.34). □

7.9 Appendix: Proofs for convex sequences

Proof of Theorem 7.9. For any set T of the form (7.21), using the concentration property (7.14) of the random variable (7.13) with $K=\mathcal{S}^{\cap}_{|T|}$, we have with probability greater than $1-\alpha$,

$$|\Pi_{\mathcal{S}^{\cap}_{|T|}}(\boldsymbol{\xi}_T)|_2^2 \le 2\delta\left(\mathcal{S}^{\cap}_{|T|}\right) + 10\log(1/\alpha) \le 20\log(en) + 10\log(1/\alpha),$$

where we used (7.26) for the last inequality. There are less that n^2 sets $T\subset\{1,\dots,n\}$ of the form (7.21). Using the union bound for all sets T of the form (7.21), we have $\mathbb{P}(\Omega(\alpha))\ge 1-\alpha$ where

$$\Omega(\alpha) \coloneqq \left\{\sup_{T\in\{T_{s,e},1\le s\le e\le n\}} |\Pi_{\mathcal{S}^{\cap}_{|T|}}(\boldsymbol{\xi}_T)|_2^2 \le \sigma^2\left(20\log(en) + 10\log\left(\frac{n^2}{\alpha}\right)\right)\right\}.$$

Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ for notational simplicity. Then (7.6) with $\boldsymbol{u}$ replaced by $\boldsymbol{\mu}$ can be rewritten as

$$|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}|_2^2 \le 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) - |\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}|_2^2.$$

By definition of $q(\cdot)$, there exists a partition $(\hat{T}_1,\dots,\hat{T}_{\hat{q}})$ of $\{1,\dots,n\}$ such that $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})$ is affine on each $\hat{T}_j$, $j=1,\dots,\hat{q}$. Furthermore, each $\hat{T}_j$ has the form (7.21) because $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\cup})\in\mathcal{S}_n^{\cup}$. We have

$$\begin{aligned}
2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}) - |\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}|_2^2 &= \sum_{j=1}^{\hat{q}} 2\boldsymbol{\xi}_{\hat{T}_j}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})_{\hat{T}_j} - |(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})_{T_j}|_2^2, \\
&\le \sum_{j=1}^{\hat{q}} \left(\frac{\boldsymbol{\xi}_{\hat{T}_j}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})_{\hat{T}_j}}{|(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})_{\hat{T}_j}|_2}\right)^2,
\end{aligned}$$

where we have used $2ab-a^2\le b^2$. By definition of $(\hat{T}_1,\dots,\hat{T}_{\hat{q}})$, $\hat{\boldsymbol{\mu}}$ is affine on $\hat{T}_j$ for each $j=1,\dots,\hat{q}$, thus the vector $(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})_{\hat{T}_j}\in\mathcal{S}^{\cap}_{|\hat{T}_j|}$ is a concave sequence. By taking the supremum, we obtain

$$|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}|_2^2 \le \sum_{j=1}^{\hat{q}} \sup_{\boldsymbol{v}\in\mathcal{S}^{\cap}_{|\hat{T}_j|}:|\boldsymbol{v}|_2^2\le 1} (\boldsymbol{\xi}_{\hat{T}_j}^T\boldsymbol{v})^2 = \sum_{j=1}^{\hat{q}} |\Pi_{\mathcal{S}^{\cap}_{|\hat{T}_j|}}(\boldsymbol{\xi}_{|\hat{T}_j|})|_2^2,$$

where we used (7.13) for the last equality. On the event $\Omega(\alpha)$ and by definition of $\hat{r}_{\cup}$,

$$|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}|_2^2 \le \sum_{j=1}^{\hat{q}} |\Pi_{\mathcal{S}^{\cap}_{|\hat{T}_j|}}(\boldsymbol{\xi}_{|\hat{T}_j|})|_2^2 \le \sigma^2\hat{q}\left(20\log(en) + 10\log(n^2/\alpha)\right) = n\hat{r}_{\cup}.$$

□

Bibliography

[1] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Inf. Inference*, 3(3):224–294, 2014. ISSN 2049-8764. doi: 10.1093/imaiai/iau005. URL http://dx.doi.org/10.1093/imaiai/iau005.

[2] Sylvain Arlot and Francis R Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems*, pages 46–54, 2009.

[3] Jean-Yves Audibert. No fast exponential deviation inequalities for the progressive mixture rule. *arXiv preprint math/0703848*, 2007.

[4] Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 2009. ISSN 0090-5364. doi: 10.1214/08-AOS623. URL http://dx.doi.org/10.1214/08-AOS623.

[5] Fadoua Balabdaoui and Jon A. Wellner. Estimation of a k-monotone density: limit distribution theory and the spline connection. *Ann. Statist.*, 35(6): 2536–2564, 2007. ISSN 0090-5364. doi: 10.1214/009053607000000262. URL http://dx.doi.org/10.1214/009053607000000262.

[6] Fadoua Balabdaoui and Jon A. Wellner. Estimation of a k -monotone density: characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1):45–70, February 2010. doi: 10.1111/j.1467-9574.2009.00438.x. URL http://dx.doi.org/10.1111/j.1467-9574.2009.00438.x.

[7] Moulinath Banerjee and Jon A. Wellner. Likelihood ratio tests for monotone functions. *Ann. Statist.*, 29(6):1699–1731, 2001. ISSN 0090-5364. doi: 10.1214/aos/1015345959. URL http://dx.doi.org/10.1214/aos/1015345959.

[8] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the Gaussian setting. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3): 1092–1119, 2014. ISSN 0246-0203. doi: 10.1214/13-AIHP539. URL http://dx.doi.org/10.1214/13-AIHP539.

[9] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006. ISSN 0178-8051. doi: 10.1007/s00440-005-0462-3. URL http://dx.doi.org/10.1007/s00440-005-0462-3.

[10] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005. ISSN 0090-5364. doi: 10.1214/009053605000000282. URL http://dx.doi.org/10.1214/009053605000000282.

[11] Pierre C. Bellec. Optimal bounds for aggregation of affine estimators. *arXiv:1410.0346, Submitted*, 2014. URL http://arxiv.org/abs/1410.0346.

[12] Pierre C. Bellec. Optimal exponential bounds for aggregation of density estimators. *Accepted in Bernoulli*, 2014. URL http://arxiv.org/abs/1405.3907.

[13] Pierre C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *arXiv:1510.08029, Submitted*, 2015. URL http://arxiv.org/abs/1510.08029.

[14] Pierre C. Bellec and Alexandre B. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.*, 16:1879–1892, 2015. ISSN 1532-4435.

[15] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. ISSN 1350-7265. doi: 10.3150/11-BEJ410. URL http://dx.doi.org/10.3150/11-BEJ410.

[16] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root Lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1204. URL http://dx.doi.org/10.1214/14-AOS1204.

[17] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. ISSN 0090-5364. doi: 10.1214/08-AOS620. URL http://dx.doi.org/10.1214/08-AOS620.

[18] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. ISSN 1435-9855. doi: 10.1007/s100970100031. URL http://dx.doi.org/10.1007/s100970100031.

[19] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007. ISSN 0178-8051. doi: 10.1007/s00440-006-0011-8. URL http://dx.doi.org/10.1007/s00440-006-0011-8.

[20] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[21] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

[22] T. Tony Cai and Mark G. Low. Adaptive confidence balls. *Ann. Statist.*, 34 (1):202–228, 2006. ISSN 0090-5364. doi: 10.1214/009053606000000146. URL http://dx.doi.org/10.1214/009053606000000146.

[23] T. Tony Cai, Mark G. Low, and Yin Xia. Adaptive confidence intervals for regression functions under shape constraints. *Ann. Statist.*, 41(2):722–750, 2013. ISSN 0090-5364. doi: 10.1214/12-AOS1068. URL http://dx.doi.org/10.1214/12-AOS1068.

[24] Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001523. URL http://dx.doi.org/10.1214/009053606000001523.

[25] Olivier Catoni. Statistical learning theory and stochastic optimization, lectures on probability theory and statistics, ecole d'eté de probabilités de saint-flour xxxi–2001. *Lecture Notes in Mathematics*, 1851:1–269, 2004.

[26] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12 (6):805–849, 2012. ISSN 1615-3375. doi: 10.1007/s10208-012-9135-7. URL http://dx.doi.org/10.1007/s10208-012-9135-7.

[27] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800, 2015. ISSN 0090-5364. doi: 10.1214/15-AOS1324. URL http://dx.doi.org/10.1214/15-AOS1324.

[28] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On matrix estimation under monotonicity constraints. *arXiv preprint arXiv:1506.03430*, 2015.

[29] Sourav Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1254. URL http://dx.doi.org/10.1214/14-AOS1254.

[30] Xi Chen, Qihang Lin, and Bodhisattva Sen. On degrees of freedom of projection estimators with applications to multivariate shape restricted regression. *arXiv preprint arXiv:1509.01877*, 2015.

[31] Arthur Cohen. All admissible linear estimates of the mean vector. *Ann. Math. Statist.*, 37:458–463, 1966. ISSN 0003-4851.

[32] Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.*, 40(3):1878–1905, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1025. URL http://dx.doi.org/10.1214/12-AOS1025.

[33] Dong Dai, Philippe Rigollet, Lucy Xia, and Tong Zhang. Aggregation of affine estimators. *Electron. J. Statist.*, 8(1):302–327, 2014. doi: 10.1214/14-ejs886. URL http://dx.doi.org/10.1214/14-ejs886.

[34] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012. ISSN 0022-0000. doi: 10.1016/j.jcss.2011.12.023. URL http://dx.doi.org/10.1016/j.jcss.2011.12.023.

[35] Arnak S. Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1038. URL http://dx.doi.org/10.1214/12-AOS1038.

[36] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007. doi: 10.1007/978-3-540-72927-3_9. URL http://dx.doi.org/10.1007/978-3-540-72927-3_9.

[37] Arnak S. Dalalyan and Alexandre B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012. ISSN 1350-7265. doi: 10.3150/11-BEJ361. URL http://dx.doi.org/10.3150/11-BEJ361.

[38] Arnak S Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *arXiv preprint arXiv:1402.1700*, 2014.

[39] Holger Dette, Axel Munk, and Thorsten Wagner. Estimating the variance in nonparametric regression—what is a reasonable choice? *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(4):751–764, 1998. ISSN 1369-7412. doi: 10.1111/1467-9868.00152. URL http://dx.doi.org/10.1111/1467-9868.00152.

[40] David L. Donoho, Richard C. Liu, and Brenda MacGibbon. Minimax risk over hyperrectangles, and implications. *Ann. Statist.*, 18(3):1416–1437, 1990. ISSN 0090-5364. doi: 10.1214/aos/1176347758. URL http://dx.doi.org/10.1214/aos/1176347758.

[41] David L. Donoho, Iain M. Johnstone, Jeffrey C. Hoch, and Alan S. Stern. Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B*, 54(1):41–81, 1992. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1992)54:1<41:MEATNB>2.0.CO;2-I&origin=MSN. With discussion and a reply by the authors.

[42] Lutz Dümbgen. Optimal confidence bands for shape-restricted curves. *Bernoulli*, 9(3):423–449, 2003. ISSN 1350-7265. doi: 10.3150/bj/1065444812. URL http://dx.doi.org/10.3150/bj/1065444812.

[43] S. Yu. Efroĭmovich and M. S. Pinsker. A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.*, (11):58–65, 1984.

[44] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. ISSN 0090-5364. doi: 10.1214/009053604000000067. URL http://dx.doi.org/10.1214/009053604000000067. With discussion, and a rejoinder by the authors.

[45] Fuchang Gao and Jon A. Wellner. Entropy estimate for high-dimensional monotonic functions. *J. Multivariate Anal.*, 98(9):1751–1764, 2007. ISSN 0047-259X. doi: 10.1016/j.jmva.2006.09.003. URL http://dx.doi.org/10.1016/j.jmva.2006.09.003.

[46] Sébastien Gerchinovitz. *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation*

techniques. PhD thesis, Université Paris Sud-Paris XI, 2011. URL https://tel.archives-ouvertes.fr/tel-00653550.

[47] Christophe Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008. ISSN 1350-7265. doi: 10.3150/08-BEJ135. URL http://dx.doi.org/10.3150/08-BEJ135.

[48] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015. ISBN 978-1-4822-3794-8.

[49] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518, 2012. ISSN 0883-4237. doi: 10.1214/12-STS398. URL http://dx.doi.org/10.1214/12-STS398.

[50] Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields*, 163 (1-2):379–411, 2015. ISSN 0178-8051. doi: 10.1007/s00440-014-0595-3. URL http://dx.doi.org/10.1007/s00440-014-0595-3.

[51] Peter Hall, JW Kay, and DM Titterinton. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3): 521–528, 1990.

[52] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42:1079–1083, 1971. ISSN 0003-4851.

[53] Mohamed Hebiri and Johannes Lederer. How correlations influence lasso prediction. *IEEE Trans. Inform. Theory*, 59(3):1846–1854, March 2013. doi: 10.1109/tit.2012.2227680. URL http://dx.doi.org/10.1109/tit.2012.2227680.

[54] C Heil. *A Basis Theory Primer: Expanded Edition*. Birkhäuser/Springer, Boston, 2011.

[55] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 6, 2012. ISSN 1083-589X. doi: 10.1214/ECP.v17-2079. URL http://dx.doi.org/10.1214/ECP.v17-2079.

[56] John Immerkær. Fast noise variance estimation. *Computer Vision and Image Understanding*, 64(2):300–302, September 1996. doi: 10.1006/cviu.1996.0060. URL http://dx.doi.org/10.1006/cviu.1996.0060.

[57] I. M. Johnstone. Function estimation and gaussian sequence models. *Unpublished manuscript*, 2(5.3):2, 2002.

[58] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008. ISSN 0090-5364. doi: 10.1214/07-AOS546. URL http://dx.doi.org/10.1214/07-AOS546.

[59] G. Kerkyacharian, A. B. Tsybakov, V. Temlyakov, D. Picard, and V. Koltchinskii. Optimal exponential bounds on the accuracy of classification. *Constr. Approx.*, 39(3):421–444, 2014. ISSN 0176-4276. doi: 10.1007/s00365-014-9229-3. URL http://dx.doi.org/10.1007/s00365-014-9229-3.

[60] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006. ISSN 0090-5364. doi: 10.1214/009053606000001019. URL http://dx.doi.org/10.1214/009053606000001019.

[61] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011. ISSN 0090-5364. doi: 10.1214/11-AOS894. URL http://dx.doi.org/10.1214/11-AOS894.

[62] Rafał Latała. Tail and moment estimates for some types of chaos. *Studia Math.*, 135(1):39–53, 1999. ISSN 0039-3223.

[63] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000. ISSN 0090-5364. doi: 10.1214/aos/1015957395. URL http://dx.doi.org/10.1214/aos/1015957395.

[64] Guillaume Lecué. Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.*, 7:971–981, 2006. ISSN 1532-4435.

[65] Guillaume Lecué. Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli*, 19(5B):2153–2166, 2013. ISSN 1350-7265. doi: 10.3150/12-BEJ447. URL http://dx.doi.org/10.3150/12-BEJ447.

[66] Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145(3-4):591–613, 2009. ISSN 0178-8051. doi: 10.1007/s00440-008-0180-8. URL http://dx.doi.org/10.1007/s00440-008-0180-8.

[67] Guillaume Lecué and Shahar Mendelson. Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli*, 16(3):605–613, 2010. ISSN 1350-7265. doi: 10.3150/09-BEJ225. URL http://dx.doi.org/10.3150/09-BEJ225.

[68] Guillaume Lecué and Shahar Mendelson. On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli*, 19(2):646–675, 2013. ISSN 1350-7265. doi: 10.3150/11-BEJ408. URL http://dx.doi.org/10.3150/11-BEJ408.

[69] Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *Ann. Statist.*, 42(1):211–224, 2014. ISSN 0090-5364. doi: 10.1214/13-AOS1190. URL http://dx.doi.org/10.1214/13-AOS1190.

[70] Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.878172. URL http://dx.doi.org/10.1109/TIT.2006.878172.

[71] K. Lounici. Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.*, 16(3):246–259, 2007. ISSN 1066-5307. doi: 10.3103/S1066530707030040. URL http://dx.doi.org/10.3103/S1066530707030040.

[72] Julien Mairal and Bin Yu. Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*, 2012.

[73] Colin L Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973.

[74] Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 1997. ISSN 0090-5364. doi: 10.1214/aos/1034276635. URL http://dx.doi.org/10.1214/aos/1034276635.

[75] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[76] Jiri Matousek and Jan Vondrak. The probabilistic method, lecture notes. 2008. URL http://kam.mff.cuni.cz/~matousek/.

[77] Michael B. McCoy and Joel A. Tropp. From Steiner formulas for cones to concentration of intrinsic volumes. *Discrete Comput. Geom.*, 51(4):926–963, 2014. ISSN 0179-5376. doi: 10.1007/s00454-014-9595-4. URL http://dx.doi.org/10.1007/s00454-014-9595-4.

[78] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009. ISSN 0090-5364. doi: 10.1214/07-AOS582. URL http://dx.doi.org/10.1214/07-AOS582.

[79] Shahar Mendelson. On aggregation for heavy-tailed classes. *arXiv preprint arXiv:1502.07097*, 2015.

[80] Mary Meyer and Michael Woodroofe. On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, 28(4):1083–1104, 2000. ISSN 0090-5364. doi: 10.1214/aos/1015956708. URL http://dx.doi.org/10.1214/aos/1015956708.

[81] Axel Munk, Nicolai Bissantz, Thorsten Wagner, and Gudrun Freitag. On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):19–41, 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00486.x. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00486.x.

[82] A.M. Nemirovski, B.T. Polyak, and Tsybakov A.B. Rate of convergence of nonparametric estimators of maximum-likelihood type. *Problems of Information Transmission*, 21:258–272, 1985.

[83] Arkadi Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000.

[84] Samet Oymak and Babak Hassibi. Sharp mse bounds for proximal denoising. *Found Comput Math*, October 2015. doi: 10.1007/s10208-015-9278-4. URL http://dx.doi.org/10.1007/s10208-015-9278-4.

[85] Vu Pham, Laurent El Ghaoui, and Arturo Fernandez. Robust sketching for multiple square-root lasso problems. *arXiv preprint arXiv:1411.0024*, 2014.

[86] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[87] A. Rakhlin, K. Sridharan, and A.B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *arXiv:1308.1147*, 2013. To appear in Bernoulli.

[88] Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *arXiv preprint arXiv:1308.1147*, 2013.

[89] Ph. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007. ISSN 1066-5307. doi: 10.3103/S1066530707030052. URL http://dx.doi.org/10.3103/S1066530707030052.

[90] Philippe Rigollet. Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 2012. ISSN 0090-5364. doi: 10.1214/11-AOS961. URL http://dx.doi.org/10.1214/11-AOS961.

[91] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011. ISSN 0090-5364. doi: 10.1214/10-AOS854. URL http://dx.doi.org/10.1214/10-AOS854.

[92] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011. ISSN 0090-5364. doi: 10.1214/10-AOS854. URL http://dx.doi.org/10.1214/10-AOS854.

[93] Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012. ISSN 0883-4237. doi: 10.1214/12-STS393. URL http://dx.doi.org/10.1214/12-STS393.

[94] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, 18:no. 82, 9, 2013. ISSN 1083-589X. doi: 10.1214/ECP.v18-2865. URL http://dx.doi.org/10.1214/ECP.v18-2865.

[95] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. ISSN 0006-3444. doi: 10.1093/biomet/ass043. URL http://dx.doi.org/10.1093/biomet/ass043.

[96] A.B. Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians*, Seoul, 2014. To appear.

[97] A.B. Tsybakov. Aggregation and minimax optimality in high dimensional estimation. *Proceedings of International Congress of Mathematicians (Seoul, 2014)*, 3:225–246, 2014.

[98] Alexandre B. Tsybakov. chapter Optimal Rates of Aggregation, pages 303–313. Springer Science + Business Media, 2003. doi: 10.1007/978-3-540-45167-9_23. URL http://dx.doi.org/10.1007/978-3-540-45167-9_23.

[99] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794. URL http://dx.doi.org/10.1007/b13794. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[100] Sara van de Geer and Johannes Lederer. The Lasso, correlated design, and improved oracle inequalities. In *From probability to statistics and back: high-dimensional models and processes*, volume 9 of *Inst. Math. Stat. (IMS) Collect.*, pages 303–316. Inst. Math. Statist., Beachwood, OH, 2013. doi: 10.1214/12-IMSCOLL922. URL http://dx.doi.org/10.1214/12-IMSCOLL922.

[101] V. N. Vapnik and A. Ya. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya*. Izdat. "Nauka", Moscow, 1974.

[102] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[103] Roman Vershynin. Estimation in high dimensions: a geometric perspective. *arXiv preprint arXiv:1405.5103*, 2014.

[104] Marten H. Wegkamp. Quasi-universal bandwidth selection for kernel density estimators. *Canad. J. Statist.*, 27(2):409–420, 1999. ISSN 0319-5724. doi: 10.2307/3315649. URL http://dx.doi.org/10.2307/3315649.

[105] F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *Ann. Probability*, 1(6):1068–1070, 1973.

[106] Yuhong Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28 (1):75–87, 2000. ISSN 0090-5364. doi: 10.1214/aos/1016120365. URL http://dx.doi.org/10.1214/aos/1016120365.

[107] Cun-Hui Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2): 528–555, 2002. ISSN 0090-5364. doi: 10.1214/aos/1021379864. URL http://dx.doi.org/10.1214/aos/1021379864.

Chapter 8

Résumé substantiel

8.1 Bornes optimales pour l'agrégation d'estimateurs affines

Nous étudions le problème d'estimation un vecteur inconnu $\boldsymbol{f} = (f_1, ..., f_n)^T \in \mathbb{R}^n$ a partir d'observations bruitées

$$Y_i = f_i + \xi_i, \qquad i = 1, ..., n, \tag{8.1}$$

où les variables aléatoires $\xi_1, ..., \xi_n$ sont i.i.d. $\mathcal{N}(0, \sigma^2)$ et représentent le bruit: La qualité d'estimation d'un estimateur du vecteur inconnue $\boldsymbol{f}$ est donnée par la norme euclidienne au carré:

$$\|\boldsymbol{f} - \hat{\boldsymbol{\mu}}\|_2^2,$$

étant donné un estimateur $\hat{\boldsymbol{\mu}}$ de $\boldsymbol{f}$. Lorsque les variables aléatoires de bruits sont des gaussiennes standards, (8.1) est le modèle des séquences gaussiennes, qui a été étudié de manière exhaustive, voir par exemple [57]. De nombreux estimateurs ont été proposés pour estimer le vecteur inconnu $\boldsymbol{f}$ à partir des observations : l'estimateur des moindres carrés, l'estimateur de Ridge, l'estimateur de Stein et les méthodes basés sur le seuillage, pour en citer quelques uns. La plupart de ces estimateurs dépendent d'un paramètre qui doit être choisi avec précaution pour obtenir des bornes d'erreur satisfaisantes. Ces estimateurs ont différentes forces et faiblesses dans différents scénarios, donc il est important de pouvoir imiter la performance du meilleur estimateur dans une famille donnée, sans faire d'hypothèse sur le vecteur inconnu $\boldsymbol{f}$. Le problème de pouvoir imiter la performance du meilleur estimateur dans une famille donnée est le problème d'agrégation de type sélection de modèles, introduit dans [83, 98]. Concrètement, soit $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ des estimateurs de $\boldsymbol{f}$ construits a partir des observations $\mathbf{y} = (Y_1, ..., Y_n)^T$. Le but est de construire avec les mêmes données (le vecteur d'observations $\mathbf{y} = (Y_1, ..., Y_n)^T$) un nouvel estimateur $\hat{\boldsymbol{\mu}}$ appelé "l'agrégat", qui satisfasse avec probabilité plus grande que $1 - \delta$ l'inégalité d'oracle exacte [1]

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + \textsc{price}_M(\delta), \tag{8.2}$$

où $\textsc{price}_M(\cdot)$ est une fonction de δ qui doit être petite. Le terme $\textsc{price}_M(\cdot)$ représente le prix à payer pour l'agrégation des estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$. Si les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ sont des vecteurs déterministes, le prix à payer pour l'agrégation de ces

[1] Par exacte, nous entendons que la constante devant le terme $\min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2$ est 1.

estimateurs est de l'ordre de $\sigma^2 \log(M/\delta)$ et (8.2) est satisfaite par un estimateur $\hat{\boldsymbol{\mu}}$ basé sur la Q-agrégation [32]. Il est intéressant de considérer des estimateurs déterministes si deux échantillons indépendants sont disponibles, de sorte que les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ sont construits a partir du premier échantillon, et l'agrégation de ces estimateurs est réalisée en utilisant le second échantillon. Dans cette situation, le premier échantillon et les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ peuvent être considérés comme gelés pour la phase d'agrégation (pour plus de détails, cf. [96]). Si les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ sont aléatoires et dépendent des données $\mathbf{y}$ utilisées pour la phase d'agrégation, deux questions naturelles se posent.

1. Est-ce que le prix à payer pour agréger les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ est plus fort lorsque les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ et les données $\mathbf{y}$ utilisées pour la phase d'agrégation ne sont pas indépendants ? Ou bien, ce prix est toujours de l'ordre de $\sigma^2 \log(M/\delta)$? Y a-t-il un prix supplémentaire a payer pour prendre en compte la dépendance ?

2. Une quantité naturelle qui capture la complexité statistique d'un estimateur $\hat{\boldsymbol{\mu}}_j$ est sa variance, définie par $\mathbb{E}|\hat{\boldsymbol{\mu}}_j - \mathbb{E}\hat{\boldsymbol{\mu}}_j|_2^2$. Si les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ sont déterministes, leurs variances sont toutes égales a 0. Maintenant que les estimateurs sont aléatoires, est-ce que le prix a payer pour les agréger va dépendre de leur complexité statistique, par exemple a travers leurs variances ? Est-il plus difficile d'agréger des estimateurs ayant une grande variance ?

Le but du présent article est de répondre a ces questions pour les estimateurs affines.

Parmi les procédures existantes pour estimer le vecteur inconnue $\boldsymbol{f}$, plusieurs sont linéaires par rapport aux observations $Y_1, ..., Y_n$. C'est le cas de l'estimateur des moindres carrés et de l'estimateur de Ridge, tandis que les estimateurs base sur le seuillage sont des fonctions non linéaires des observations. Des exemples d'estimateurs linéaires ou affines par rapport aux observations sont donnés dans [35, Section 1.2], [2]. Un estimateur affine est de la forme $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j$, où A_j est une matrice déterministe de taille $n \times n$ et $\boldsymbol{b}_j \in \mathbb{R}^n$ est un vecteur déterministe dans $\mathbb{R}^n$. La linéarité des estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ permet de traiter explicitement la dépendance entre les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ et les données $\mathbf{y} = (Y_1, ..., Y_n)^T$ utilisées pour la phase d'agrégation.

Les articles [70, 35, 33] ont étudié différentes procédures qui satisfont des inégalités d'oracles exactes pour le problème d'agrégation d'estimateurs affines lorsque le bruit est gaussien. Leung and Barron [70], Dalalyan and Salmon [35] ont proposé un estimateur $\hat{\boldsymbol{\mu}}^{EW}$ basé sur les poids exponentiels pour lequel l'inégalité d'oracle suivante est vérifiée en espérance :

$$\mathbb{E}|\boldsymbol{f} - \hat{\boldsymbol{\mu}}^{EW}|_2^2 \leq \min_{j=1,...,M} \mathbb{E}|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 8\sigma^2 \log M,$$

sous l'hypothèse que tous les A_j sont des projecteurs orthogonaux, ou sous une hypothèse forte de commutativité des matrices A_j. La constante 8 peut être réduite a 4 si toutes les matrices A_j sont des projecteurs orthogonaux. Si les matrices A_j ne sont pas symétriques, [35] a montrée qu'il était possible de symétriser les estimateurs affines avant la phase d'agrégation obtenant une inégalité d'oracle similaire, ce qui suggère que l'hypothèse de symétrie n'est pas nécessaire. Bien que l'estimateur $\hat{\boldsymbol{\mu}}^{EW}$ satisfasse l'inégalité ci-dessus en espérance, il a été montré dans [3, 32] que cet estimateur ne peut pas satisfaire une inégalité similaire avec grande probabilité, avec

un terme d'erreur inévitable de l'ordre de $\sqrt{n}$. Dans Dai et al. [33], une inégalité d'oracle exacte est démontré pour un estimateur $\hat{\boldsymbol{\mu}}^Q$ basé sur la Q-aggregation [90, 32]. Plus précisément, [33] a montré que si les matrices $A_1, ..., A_M$ sont symétriques et positives semi-définies, alors l'estimateur $\hat{\boldsymbol{\mu}}^Q$ verifie avec probabilité plus grande que $1 - \delta$

$$|\boldsymbol{f} - \hat{\boldsymbol{\mu}}^Q|_2^2 \leq \min_{j=1,...,M} \left(|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 4\sigma^2 \mathrm{Tr}(A_j) \right) + C\sigma^2 \log(M/\delta), \tag{8.3}$$

où la constante C est proportionnelle à la plus grande norme opérateur des matrices $A_1, ..., A_M$. Le terme $4\sigma^2 \mathrm{Tr}(A_j)$ est intimement lié à la complexité statistique de l'estimateur $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j$. En effet, la variance de l'estimateur $\hat{\boldsymbol{\mu}}_j$ est $\mathbb{E}|\hat{\boldsymbol{\mu}}_j - \mathbb{E}\hat{\boldsymbol{\mu}}_j|_2^2 = \sigma^2 \mathrm{Tr}(A_j^T A_j)$. Si $\hat{\boldsymbol{\mu}}_j$ est un estimateur des moindres carrés, A_j est un projecteur orthogonal et sa variance est $\sigma^2 \mathrm{Tr} A_j$. La complexité statistique de l'estimateur $\hat{\boldsymbol{\mu}}_j$ apparait clairement dans le terme résiduel de l'inégalité d'oracle (8.3) prouvé dans [33]. On pourrait donc penser que le prix à payer pour agréger M estimateurs affines, i.e. la fonction $\textsc{price}_M(\delta)$ dans (8.2), dépend de la complexité statistique des estimateurs a agréger. La borne (8.3) pourrait amener à la conclusion que le prix à payer pour l'agrégation d'estimateurs affines peut être substantiellement plus grand que $\sigma^2 \log(M/\delta)$ qui est le prix a payer pour agréger des vecteurs déterministes. En effet, le terme supplémentaire $4\sigma^2 \mathrm{Tr}(A_j)$ peut être large dans des situations où la trace de certaines matrices A_j est grande. Par exemple, si l'on agrège les estimateurs $\hat{\boldsymbol{\mu}}_1 = \lambda_1 \mathbf{y}, ..., \hat{\boldsymbol{\mu}}_M = \lambda_M \mathbf{y}$, où $\lambda_1, ..., \lambda_M$ sont des constantes positives, alors le terme $4\sigma^2 \mathrm{Tr}(A_j)$ présent dans l'inégalité d'oracle est de l'ordre de $\sigma^2 n \lambda_j$, donc ce terme peut être plus grande que la vitesse optimale d'agrégation $\sigma^2 \log M$. Ce terme $4\sigma^2 \mathrm{Tr}(A_j)$ rend l'inégalité d'oracle (8.3) intéressante seulement pour des scénarios où les matrices A_j ont une trace faible. Mais pour en revenir à la question plus fondamentale évoquée plus haut, le terme $\sigma^2 \mathrm{Tr} A_j$ suggère que le prix à payer pour agréger les estimateurs affines augmente avec la complexité statistique des estimateurs à agréger. Enfin, les résultats discutés plus haut nécessitent des hypothèses spécifiques sur les matrices $A_1, ..., A_M$ [70, 35, 33]. Cela pose une troisième question :

3. Est-ce que la nature des matrices $A_1, ..., A_m$ a un impact sur le prix à payer pour agréger les estimateurs affines correspondants ? Est-ce que le prix (8.2) est plus faible lorsque ces matrices sont des projecteurs orthogonaux, ou bien lorsqu'elle sont symétriques définies positives ?

La contribution principale du présent article est de répondre aux trois questions posées plus haut :

1. Théorème 8.1 montre que la minimisation d'un critère de moindres carrés pénalisé sur le simplex permet de construire un estimateur qui vérifie l'inégalité d'oracle exacte (8.2) avec $\textsc{price}_M(\delta) = c\sigma^2 \log(M/\delta)$ où $c > 0$ est une constante absolue. Ce prix est du même ordre que le prix à payer pour l'agrégation de vecteurs déterministes. La dépendance entre les estimateurs à agréger et les données $\mathbf{y}$ n'induit donc pas de terme d'erreur supplémentaire.

2. La forme des estimateurs affines à agréger n'a pas d'impact sur le prix à payer pour les agréger. En particulier, les inégalités d'oracle exactes du présent article ne font pas intervenir de quantités dépendant des matrices A_j tels que $\sigma^2 \mathrm{Tr} A_j$.

3. La seule hypothèse faite sur les matrices $A_1, ..., A_M$ est que $|||A_j|||_2 \leq 1$ pour tout $j = 1, ..., M$, où $|||\cdot|||_2$ est la norme opérateur. Toutes les autres hypothèse

sur les matrices $A_1, ..., A_M$ ne sont pas nécessaires, en particulier les matrices peuvent ne pas être symétriques ou avoir des valeurs propres négatives.

L'organisation de l'article est la suivante. La Section 8.1.1 définit les notations utilisées dans l'article. La Section 8.1.2 définit un estimateur par la minimisation d'un critère pénalisé sur le simplex, et montre que que cet estimateur vérifie une inégalité d'oracle exacte en déviation pour l'agrégation d'estimateurs affines. Le role de la pénalité est étudié dans la Section 8.1.3 et la Section 8.1.4. Dans la Section 8.1.5 nous considérons des poids a priori sur les estimateurs à agréger.

8.1.1 Notation

Soit $\boldsymbol{f} = (f_1, ..., f_n)^T \in \mathbb{R}^n$ un vecteur de régression inconnue. Nous observons n variables aléatoires (8.1) où $\xi_1, ..., \xi_n$ sont des variables aléatoires sous-gaussiennes telles $\mathbb{E}[\xi_i] = 0$ et $\mathbb{E}[\xi_i^2] = \sigma^2$. Avec la notation vectorielle, nous avons $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\xi}$ où $\mathbf{y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{f} = (f_1, ..., f_n)^T$ et $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^T$.

Étant donné un estimateur $\hat{\boldsymbol{\mu}}$ de $\boldsymbol{f}$, nous mesurons l'erreur d'estimation avec la perte quadratique $|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2$, où $|\cdot|_2$ est la norme euclidienne dans $\mathbb{R}^n$. Soit $M \geq 2$ un entier. Nous considérons M estimateurs affines de la forme

$$\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j, \qquad j = 1, ..., M.$$

Les matrices $A_1, ..., A_M$ et les vecteurs $\boldsymbol{b}_1, ..., \boldsymbol{b}_M \in \mathbb{R}^n$ sont déterministes. Nous définissons le simplex dans $\mathbb{R}^M$ par :

$$\Lambda^M = \left\{ \boldsymbol{\theta} \in \mathbb{R}^M, \quad \sum_{j=1}^M \theta_j = 1, \quad \forall j = 1 \dots M, \ \ \theta_j \geq 0 \right\}.$$

Pour tout $\boldsymbol{\theta} \in \Lambda^M$, soit

$$A_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j A_j, \qquad \boldsymbol{b}_{\boldsymbol{\theta}} = \sum_{j=1}^M \theta_j \boldsymbol{b}_j, \qquad \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = A_{\boldsymbol{\theta}} \mathbf{y} + \boldsymbol{b}_{\boldsymbol{\theta}}.$$

Soit $\boldsymbol{e}_1, ..., \boldsymbol{e}_M$ les vecteurs de la base canonique $\mathbb{R}^M$. Avec ces notations, $\hat{\boldsymbol{\mu}}_j = \hat{\boldsymbol{\mu}}_{\boldsymbol{e}_j}$ pour tout $j = 1, ..., M$.

Un projecteur orthogonal est une matrice P de taille $n \times n$ telle que

$$P = P^T = P^2.$$

Soit $I_{n \times n}$ la matrice identité de taille $n \times n$. Étant donnée une matrice réelle $A = (a_{i,j})_{i,j=1,...,n}$ de taille $n \times n$, la norme opérateur de A, la norme de Frobenius de A et la norme nucléaire de A sont respectivement définies par :

$$|||A|||_2 = \sup_{x \neq 0} \frac{|Ax|_2}{|x|_2}, \quad \|A\|_{\mathrm{F}} = \sqrt{\sum_{i,j=1,...,n} a_{i,j}^2}, \quad \|A\|_1 = \mathrm{Tr}\left(\sqrt{A^T A}\right).$$

L'inégalité suivante est vérifiée pour toutes matrices carrées M, M' :

$$|||MM'|||_2 \leq |||M|||_2 |||M'|||_2, \qquad \|MM'\|_{\mathrm{F}} \leq |||M|||_2 \, \|M'\|_{\mathrm{F}}.$$

Enfin, la fonction log est le logarithme naturel vérifiant $\log(e) = 1$.

8.1.2 Une procédure pénalisée sur le simplex

Pour tout $\boldsymbol{\theta} \in \Lambda^M$, soit

$$C_p(\boldsymbol{\theta}) := |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 - 2\mathbf{y}^T\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + 2\sigma^2\mathrm{Tr}(A_{\boldsymbol{\theta}}), \tag{8.4}$$

qui est le critère C_p de Mallows [73]. Nous définissons ensuite

$$H_{\mathrm{pen}}(\boldsymbol{\theta}) = C_p(\boldsymbol{\theta}) + \tfrac{1}{2}\mathrm{pen}(\boldsymbol{\theta}), \tag{8.5}$$

où

$$\mathrm{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{M} \theta_j |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \hat{\boldsymbol{\mu}}_j|_2^2. \tag{8.6}$$

Nous considérons l'estimateur $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}}$ solution du problème de minimisation

$$\hat{\boldsymbol{\theta}}_{\mathrm{pen}} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\Lambda^M} H_{\mathrm{pen}}(\boldsymbol{\theta}). \tag{8.7}$$

La fonction H_{pen} est quadratique et convexe. Minimiser H_{pen} sur le simplex est un programme quadratique pour lequel des algorithmes efficaces sont disponibles. La convexité de H_{pen} montre que $\hat{\boldsymbol{\theta}}_{\mathrm{pen}}$ est bien défini, bien qu'il puisse ne pas être unique (par exemple, si tous les $\hat{\boldsymbol{\mu}}_j$ sont égaux alors H_{pen} est constante sur le simplex).

Nous expliquons maintenant la signification de chacun des termes qui apparaissent dans (8.5). Si $\boldsymbol{\theta}$ est fixé, alors $C_p(\boldsymbol{\theta})$ est un estimateur non biaisé de

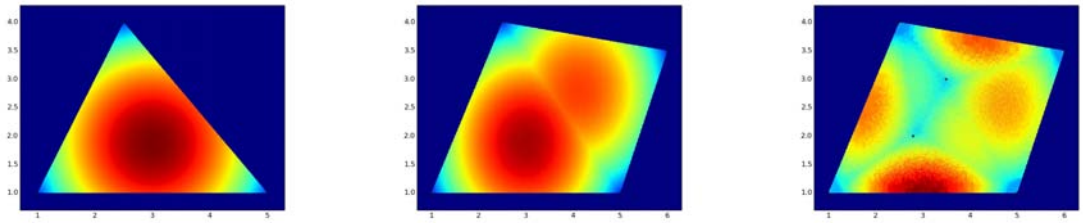
$$R(\boldsymbol{\theta}) := |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 - 2\boldsymbol{f}^T\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{f}|_2^2 - |\boldsymbol{f}|_2^2, \tag{8.8}$$

qui est égale, à une constante additive près, au risque de l'estimateur $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$.

La pénalité (8.6) a pour origine la Q-agrégation, qui est une technique avérée pour obtenir des inégalités d'oracle exacte en déviation quand la perte est fortement convexe [90, 32, 69, 12]. Puisque les estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ dépendent des données $\mathbf{y}$, la pénalité (8.6) dépend également des données, ce qui n'est pas le cas si $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ sont des vecteurs déterministes comme dans [32]. Pour donner une intuition géométrique de la pénalité (8.6), soit $c \in \mathbb{R}^n$ une solution des M équations linéaires $2c^T\hat{\boldsymbol{\mu}}_j = |\hat{\boldsymbol{\mu}}_j|_2^2, j = 1, ..., M$, et supposons seulement dans le reste du présent paragraphe que cette solution existe, bien que cette hypothèse ne puisse pas être vérifiée lorsque $M > n$. Dans ce cas, la pénalité peut être réécrite

$$\mathrm{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{M} \theta_j |\hat{\boldsymbol{\mu}}_j|_2^2 - |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 = 2c^T\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}|_2^2 = |c|_2^2 - |\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - c|_2^2. \tag{8.9}$$

Nous pouvons donc écrire $\mathrm{pen}(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})$ pour une fonction g définie sur l'enveloppe convexe de $\{\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M\}$. L'équation (8.9) montre que les lignes de niveau de g sont des sphères euclidiennes centrées en c. La fonction g est positive, elle est minimale aux points extrêmes $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ puisque $g(\hat{\boldsymbol{\mu}}_j) = 0$ pour tout $j = 1, ..., M$ et g est maximale en la projection euclidienne de c sur l'enveloppe convexe de $\{\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M\}$. Intuitivement, la pénalité (8.6) repousse $\boldsymbol{\theta}$ loin du centre c vers les points extrêmes. Ici, les lignes de niveau de la fonction $\boldsymbol{\theta} \to \mathrm{pen}(\boldsymbol{\theta})$ définie sur $\mathbb{R}^M$ sont des ellipsoïdes centrés en $\boldsymbol{\theta}_c$, où $\boldsymbol{\theta}_c$ est l'unique point de $\mathbb{R}^M$ telle que $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_c} = c$. Si $M > n$ ou bien si c n'est pas bien définie, alors les lignes de niveau de $\mathrm{pen}(\cdot)$ sont plus complexes et ne peuvent pas être décrites aussi simplement.



(a) 3 points dans $\mathbf{R}^2$. (b) 4 points dans $\mathbf{R}^2$. (c) 6 points dans $\mathbf{R}^2$.

Figure 8.1: La pénalité (8.6) et ses lignes de niveaux. La plus grande pénalité est en rouge, la plus petite en bleu.

Theorem 8.1 (Résultat principal)**.** *Soit $M \geq 2$. Pour tout $j = 1, ..., M$, soit un estimateur affine $\hat{\boldsymbol{\mu}}_j = A_j \mathbf{y} + \boldsymbol{b}_j$ et supposons que $|||A_j|||_2 \leq 1$. Supposons que les variables aléatoires de bruit $\xi_1, ...\xi_n$ sont i.i.d. $\mathcal{N}(0, \sigma^2)$. Soit $\hat{\boldsymbol{\theta}}_{\text{pen}}$ l'estimateur (8.7). Alors pour tout $x > 0$, l'estimateur $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$ vérifie avec probabilité au moins $1 - \exp(-x)$,*

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 30\sigma^2(x + 2\log M). \tag{8.10}$$

De plus,

$$\mathbb{E}\left[|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2\right] \leq \mathbb{E}\left[\min_{j=1,...,M} |\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2\right] + 60\sigma^2 \log(M). \tag{8.11}$$

L'inégalité d'oracle exacte en déviation donnée par [33] présente un terme additif proportionnel à $\sigma^2 \mathrm{Tr}(A_j)$, cf. (8.3). Une amélioration du présent article est l'absence de ce terme additif qui peut être large pour des matrices A_j ayant une trace large. Notre analyse montre que les quantités $\sigma^2 \mathrm{Tr}(A_j)$ ne sont pas significatives pour le problème d'agrégation d'estimateurs affines, et le Théorème 8.1 améliore le résultat précédemment obtenue par [33].

Les hypothèses sur les matrices $A_1, ..., A_M$ ne concernent que la norme opérateur de ces matrices. Les matrices peuvent ne pas être symétriques ou avoir des valeurs propres négatives. Le résultat ci-dessus montre donc que les restrictions sur les matrices $A_1, ..., A_M$ introduites dans [70, 35, 33] ne sont pas intrinsèques au problème d'agrégation d'estimateurs affines.

Un estimateur de la forme $B_j \mathbf{y} + \boldsymbol{b}_j$ avec $|||B_j|||_2 > 1$ et $\boldsymbol{b}_j \in \mathbb{R}^n$ n'est pas admissible au sens où il existe une matrice $A_j = A_j(B_j)$ telle que

$$|||A_j|||_2 \leq 1, \qquad \mathbb{E}\left[|A_j \mathbf{y} + \boldsymbol{b}_j - \boldsymbol{f}|_2^2\right] \leq \mathbb{E}\left[|B_j \mathbf{y} + \boldsymbol{b}_j - \boldsymbol{f}|_2^2\right] \tag{8.12}$$

pour tout $\boldsymbol{f} \in \mathbb{R}^n$, cf. Cohen [31]. Soit $B_1, ..., B_M$ des matrices réelles et $\boldsymbol{b}_1, ..., \boldsymbol{b}_M \in \mathbb{R}^n$ des vecteurs déterministes. Nous définissons les matrices $A_j = A_j(B_j)$ de la manière suivante. Si $|||B_j|||_2 > 1$ alors A_j est une matrice telle que (8.12) est vérifié et si $|||B_j|||_2 \leq 1$, alors $A_j = B_j$. Grâce au Théorème 8.1, l'estimateur $\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}}$ qui agrègent les estimateurs $(A_j \mathbf{y} + \boldsymbol{b}_j)_{j=1,...,M}$ vérifie

$$\begin{aligned}
\mathbb{E}|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 &\leq \min_{j=1,...,M} \mathbb{E}\left[|A_j \mathbf{y} + \boldsymbol{b}_j - \boldsymbol{f}|_2^2\right] + 60\sigma^2 \log(M), \\
&\leq \min_{j=1,...,M} \mathbb{E}\left[|B_j \mathbf{y} + \boldsymbol{b}_j - \boldsymbol{f}|_2^2\right] + 60\sigma^2 \log(M).
\end{aligned}$$

Nous obtenons donc une inégalité d'oracle en espérance sans l'hypothèse $\max_{j=1,\dots,M} |||B_j|||_2 \leq 1$ si les estimateurs $(B_j\mathbf{y} + \boldsymbol{b}_j)_{j=1,\dots,M}$ sont préalablement transformés en $(A_j\mathbf{y} + \boldsymbol{b}_j)_{j=1,\dots,M}$ avec $|||A_j|||_2 \leq 1$.

La proposition suivante montre que les majorations du Théorème 8.1 sont optimales dans un sens minimax. Pour tout $\boldsymbol{f} \in \mathbb{R}^n$ notons $\mathbb{P}_{\boldsymbol{f}}$ la mesure de probabilité de la variable aléatoire $\mathbf{y} = \boldsymbol{f} + \boldsymbol{\xi}$. Une minoration pour l'agrégation de vecteurs déterministe a été prouvé dans [92, Theorem 5.4 with $S = 1$]. Cette minoration implique le résultat suivant.

Proposition 8.2. *Il existe des constantes absolues $c^*, C^*, p^* > 0$ telle que le résultat suivant est vérifié. Pour tout $M, n \geq C^*$, il existe des vecteurs $\boldsymbol{b}_1, \dots, \boldsymbol{b}_M \in \mathbb{R}^n$ et des projecteurs orthogonaux $A_1, \dots, A_M$ de rang 1 tels que*

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{f}\in\mathbb{R}^n} \mathbb{P}_{\boldsymbol{f}} \left(|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 - \min_{k=1,\dots,M} |\boldsymbol{b}_k - \boldsymbol{f}|_2^2 \geq c^*\sigma^2 \log(M) \right) \geq p^*, \tag{8.13}$$

$$\inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{f}\in\mathbb{R}^n} \mathbb{P}_{\boldsymbol{f}} \left(|\hat{\boldsymbol{\mu}} - \boldsymbol{f}|_2^2 - \min_{k=1,\dots,M} |A_k\mathbf{y} - \boldsymbol{f}|_2^2 \geq c^*\sigma^2 \log(M) \right) \geq p^*, \tag{8.14}$$

où la borne inférieure est prise sur tous les estimateurs $\hat{\boldsymbol{\mu}}$.

Ce résultat implique que les majorations du Théorème 8.1 sont optimales pour le problème d'agrégation d'estimateurs affines. La minoration ci-dessus peut être construite soit avec un dictionnaire de vecteur déterministes (cf. (8.13)), soit avec un dictionnaire de projecteurs orthogonaux de rang 1 (cf. (8.14)).

8.1.3 La pénalité (8.6) améliore la sélection de modèles basée sur C_p

Pour expliquer le rôle joué par la pénalité (8.6) pour le problème d'agrégation d'estimateurs affines, considérons d'abord la procédère de sélection par minimization du critère C_p . Définissons $\hat{J}$ par

$$\hat{J} \in \operatorname*{argmin}_{j=1,\dots,M} C_p(\boldsymbol{e}_j), \tag{8.15}$$

où $C_p(\cdot)$ est défini par (8.4). En utilisant l'inégalité $C_p(\boldsymbol{e}_{\hat{J}}) \leq C_p(\boldsymbol{e}_k)$ pour tout $k = 1, \dots, M$ et les définitions de $C_p(\cdot)$ et $R(\cdot)$ données dans (8.4) et (8.8), l'inégalité suivante est vérifiée presque sûrement :

$$|\hat{\boldsymbol{\mu}}_{\hat{J}} - \boldsymbol{f}|_2^2 \leq \min_{k=1,\dots,M} |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \max_{j,k=1,\dots,M} \Delta_{jk}, \tag{8.16}$$

où $\Delta_{jk} := C_p(\boldsymbol{e}_k) - C_p(\boldsymbol{e}_j) - (R(\boldsymbol{e}_k) - R(\boldsymbol{e}_j))$. Il est donc possible de montrer une inégalité d'oracle pour l'estimateur $\hat{\boldsymbol{\mu}}_{\hat{J}}$ si nous pouvons contrôlé les quantités Δ_{jk} uniformément sur toutes les paires $j, k = 1, \dots, M$. Ces quantités peuvent être réécrites

$$\Delta_{jk} = 2\boldsymbol{\xi}^T((A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k) + 2\left(\boldsymbol{\xi}^T(A_j - A_k)\boldsymbol{\xi} - \sigma^2 \mathrm{Tr}(A_j - A_k)\right). \tag{8.17}$$

Deux quantités aléatoires apparaissent dans Δ_{jk}. La première quantité est une variable aléatoire gaussienne centrée, de variance $4\sigma^2|(A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k|_2^2$. La seconde quantité est une forme quadratique en $\boldsymbol{\xi}$, et sa variance est de l'ordre de $\sigma^4 \|A_j - A_k\|_{\mathrm{F}}^2$. Ce terme quadratique est parfois appelé un chaos gaussien d'ordre 2.

La déviation de ces deux termes est caractérisée par les inégalités de concentration suivantes. Pour tout vecteur $\boldsymbol{v} \in \mathbb{R}^n$, une majoration standard de la queue gaussienne est

$$\mathbb{P}\left(\boldsymbol{v}^T\boldsymbol{\xi} > \sigma|\boldsymbol{v}|_2\sqrt{2x}\right) \leq \exp(-x), \qquad \forall x > 0. \tag{8.18}$$

Pour le chaos gaussien d'ordre 2, l'inégalité de concentration suivante est montré dans [20, Example 2.12].

Lemma 8.3. *Supposons que $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$. Alors pour toute matrice carrée B de taille n,*

$$\mathbb{P}\left(\boldsymbol{\xi}^T B\boldsymbol{\xi} - \sigma^2 \mathrm{Tr} B > 2\sigma^2 \|B\|_{\mathrm{F}} \sqrt{x} + 2\sigma^2 |||B|||_2 x\right) \leq \exp(-x), \tag{8.19}$$

où $\sigma^2 \mathrm{Tr} B = \mathbb{E}[\boldsymbol{\xi}^T B\boldsymbol{\xi}]$.

Notons $\boldsymbol{v} = 2((A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k)$ et $B = 2(A_k - A_j)$ pour étudier le comportement de la variable aléatoire Δ_{jk}. Si $|||A_j - A_k|||_2$ est petite, (8.18) et (8.19) montrent que les déviations de Δ_{jk} sont de l'ordre des deux quantités

$$\sigma|(A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k|_2, \qquad \sigma^2 \|A_j - A_k\|_{\mathrm{F}},$$

i.e., l'écart type des deux termes aléatoires présents dans Δ_{jk}. Les inégalités de concentration (8.18) et (8.19) sont connues pour être précises [62], il y a donc peu d'espoir de majorer les larges déviations de Δ_{jk} indépendamment de $\boldsymbol{f}$, A_j et A_k pour obtenir une inégalité d'oracle exacte. Cependant, il est possible de modifier l'analyse ci-dessus pour obtenir l'inégalité d'oracle suivante, qui est cependant inexacte avec une constante principale strictement plus grande que 1.

Proposition 8.4. *Il existe des constantes absolues $c, C > 0$ telles que le résultat suivant est vérifiée. Supposons que $|||A_j|||_2 \leq 1$ pour tout $j = 1, ..., M$. Soit $0 < \varepsilon < c$ et soit $\hat{J}$ l'estimateur défini par (8.15). Pour tout $x > 0$, l'estimateur $\hat{\boldsymbol{\mu}}_{\hat{J}}$ vérifie avec probabilité plus grande que $1 - 2\exp(-x)$*

$$|\hat{\boldsymbol{\mu}}_{\hat{J}} - \boldsymbol{f}|_2^2 \leq (1+\varepsilon) \min_{k=1,...,M} |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + C\sigma^2(x + 2\log M)/\varepsilon.$$

L'estimateur $\hat{\boldsymbol{\mu}}_{\hat{J}}$ ne peut pas vérifier une inégalité d'oracle exacte avec un terme d'erreur de l'ordre de $\sigma^2 \log M$, et cette faiblesse ne peut pas être réparé par une procédure de la forme $\hat{\boldsymbol{\mu}}_{\hat{K}}$ où $\hat{K}$ est un estimateur qui prend ses values dans l'ensemble discret $\{1, ..., M\}$. En effet, il est prouvé dans [46, Section 6.4.2 et Proposition 6.1] qu'il existe $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^n$ et deux projecteurs orthogonaux A_1, A_2 tels que pour n'importe quel estimateur $\hat{K}$ valué dans $\{1, 2\}$,

$$\sup_{\boldsymbol{f}\in\{\boldsymbol{f}_1,\boldsymbol{f}_2\}} \left(\mathbb{E}|A_{\hat{K}}\mathbf{y} - \boldsymbol{f}|_2^2 - \min_{j=1,2} \mathbb{E}|A_j\mathbf{y} - \boldsymbol{f}|_2^2\right) \geq \sigma^2\sqrt{n}/4,$$

dès que n est plus grand qu'une valeur absolue. Si l'on regarde de plus près la preuve de ce résultat, on trouve que

$$\sigma|(A_2 - A_1)\boldsymbol{f} + \boldsymbol{b}_2 - \boldsymbol{b}_1|_2 \geq \sigma^2\sqrt{n}, \qquad \forall \boldsymbol{f} \in \{\boldsymbol{f}_1, \boldsymbol{f}_2\},$$

où $\boldsymbol{b}_1 = \boldsymbol{b}_2 = 0$. Cette minoration d'ordre $\sqrt{n}$ est donc lié au terme gaussien de la variable aléatoire Δ_{12}, i.e., au terme $\boldsymbol{\xi}^T((A_1 - A_2)\boldsymbol{f} + \boldsymbol{b}_1 - \boldsymbol{b}_2)$, cf. (8.17).

La procédure $\hat{\boldsymbol{\mu}}_{\hat{J}}$ ne vérifie pas d'inégalité d'oracle exacte car les variances des deux termes de Δ_{jk} peuvent être larges et ne sont pas contrôlés. Le rôle de la pénalité (8.6) est exactement de contrôler les déviations de Δ_{jk}. La proposition suivante précise cette interprétation.

Proposition 8.5. *Soit $\hat{\boldsymbol{\theta}}_{\text{pen}}$ l'estimateur (8.7). Alors, presque sûrement*

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 \leq \min_{q=1,\dots,M} \left(|\hat{\boldsymbol{\mu}}_q - \boldsymbol{f}|_2^2\right) + \max_{j,k=1,\dots,M} \left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right), \tag{8.20}$$

où Δ_{jk} est définie par (8.17). De plus, pour tout $j, k = 1, \dots, M$,

$$\mathbb{E}\left[\frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right] = \frac{1}{2}|(A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k|_2^2 + \frac{\sigma^2}{2}\|A_j - A_k\|_{\text{F}}^2. \tag{8.21}$$

La preuve de (8.20) est donnée dans la Section 8.1.4 ci-dessous. Une décomposition biais-variance montre directement (8.21), puisque $\mathbb{E}[\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k] = (A_j - A_k)\boldsymbol{f} + \boldsymbol{b}_j - \boldsymbol{b}_k$ et $\mathbb{E}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k - \mathbb{E}[\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k]|_2^2 = \mathbb{E}|(A_j - A_k)\boldsymbol{\xi}|_2^2 = \sigma^2\|A_j - A_k\|_{\text{F}}^2$.

Comparé avec (8.16), le membre de droite de (8.20) présente les quantités $-\frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2$. Nous expliquerons plus bas que ces quantités apparaissent à cause de l'interaction entre la pénalité (8.6) et la convexité forte de la fonction H_{pen}.

A partir de (8.20), voici un résumé de la preuve du Théorème 8.1. En étudiant précisément la fonction génératrice des moments, nous montrerons que pour toute paire (j, k) nous avons

$$\mathbb{E}\exp\left(\lambda_0\left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right)\right) \leq 1$$

où $\lambda_0 = (30\sigma^2)^{-1}$ si le bruit $\boldsymbol{\xi}$ a pour distribution $\mathcal{N}(0, \sigma^2 I_{n\times n})$. Nous avons donc

$$\mathbb{E}\exp\left(\lambda_0 \max_{j,k=1,\dots,M}\left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right)\right) \leq M^2.$$

En utilisant l'inégalité de Jensen, nous obtenons (8.11) et en utilisant la borne de Chernoff nous obtenons (8.10). Cela explique le succès de la pénalité (8.6) pour le problème d'agrégation de type sélection de modèle : la pénalité et la convexité forte de H_{pen} font apparaitre les quantités $-\frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2$, et ces quantités sont exactement celles nécessaires pour contrôler les larges déviations des variables aléatoires Δ_{jk}.

8.1.4 Convexité forte et la pénalité (8.6)

Pour comprendre plus précisément l'interaction entre la pénalité (8.6) et la convexité forte de H_{pen}, nous donnons maintenant une preuve de (8.20).

Preuve de (8.20). Soit $k = 1, \dots, M$ un entier fixé. Le simplex Λ^M est un ensemble convexe et la fonction H_{pen} est convexe, donc nous avons

$$\nabla H_{\text{pen}}(\hat{\boldsymbol{\theta}}_{\text{pen}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}_{\text{pen}}) \geq 0,$$

cf. [21, Section 4.2.3, equation (4.21)]. L'inégalité (8.20) est une conséquence de

$$\begin{aligned}
& |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 \\
\leq\ & |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\text{pen}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \nabla H_{\text{pen}}(\hat{\boldsymbol{\theta}}_{\text{pen}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}_{\text{pen}}), && (8.22)\\
=\ & \sum_{j=1}^M \hat{\theta}_{\text{pen},j}\left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right), && (8.23)\\
\leq\ & \max_{j=1,\dots,M}\left(\Delta_{jk} - \frac{1}{2}|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2\right). && (8.24)
\end{aligned}$$

L'égalité (8.23) est obtenue par des manipulations élémentaires tandis que (8.24) est une conséquence de $\sum_{j=1}^M \hat{\theta}_{\text{pen},j} = 1$ et $\hat{\theta}_{\text{pen},j} \geq 0$ pour tout $j = 1, \dots, M$. □

Il est possible d'interpréter cet argument en terme d'interaction entre la convexité forte et la pénalité (8.6). En effet, le membre de droite de (8.22) vérifie

$$\begin{aligned} & |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \boldsymbol{f}|_2^2 - |\hat{\boldsymbol{\mu}}_k - \boldsymbol{f}|_2^2 + \nabla H_{\mathrm{pen}}(\hat{\boldsymbol{\theta}}_{\mathrm{pen}})^T(\boldsymbol{e}_k - \hat{\boldsymbol{\theta}}_{\mathrm{pen}}) \\ = & \sum_{j=1}^{M} \hat{\theta}_{\mathrm{pen},\,j}\Delta_{jk} - \frac{1}{2}\left[\mathrm{pen}(\hat{\boldsymbol{\theta}}_{\mathrm{pen}}) + |\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \hat{\boldsymbol{\mu}}_k|_2^2\right]. \end{aligned}$$

Le terme $|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \hat{\boldsymbol{\mu}}_k|_2^2$ vient de la convexité forte de la fonction H_{pen}. Par des manipulations élémentaires, nous obtenons

$$\mathrm{pen}(\hat{\boldsymbol{\theta}}_{\mathrm{pen}}) \quad + \quad \underbrace{|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\mathrm{pen}}} - \hat{\boldsymbol{\mu}}_k|_2^2}_{\substack{\text{terme provenant de} \\ \text{la convexité forte de } H_{\mathrm{pen}}}} \quad = \quad \sum_{j=1}^{M} \hat{\theta}_{\mathrm{pen},\,j} \quad \underbrace{|\hat{\boldsymbol{\mu}}_j - \hat{\boldsymbol{\mu}}_k|_2^2}_{\substack{\text{terme qui contrôle} \\ \text{les déviations de } \Delta_{jk}}}. \tag{8.25}$$

La formule (8.25) met en lumière le rôle de la pénalité (8.6): la pénalité transforme le terme quadratique provenant de la convexité forte en un terme linéaire donné par le membre de droite de (8.25).

8.1.5 Poids a priori

Nous considérons maintenant le problème d'égrégation de M estimateurs affines étant donnée une mesure de probabilité a priori $\boldsymbol{\pi} = (\pi_1, ..., \pi_M)^T$ sur l'ensemble d'indices $\{1, ..., M\}$.

Theorem 8.6. *Soit $M \geq 2$. Pour tout $j = 1, ..., M$, considérons l'estimateur $\hat{\boldsymbol{\mu}}_j = A_j\mathbf{y} + \boldsymbol{b}_j$ et supposons que $|||A_j|||_2 \leq 1$. Soit $\boldsymbol{\pi} = (\pi_1, ..., \pi_M)^T \in \Lambda^M$. Supposons que le bruit $\boldsymbol{\xi}$ a pour distribution $\mathcal{N}(0, \sigma^2 I_{n\times n})$. Soit $\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}} \in \mathrm{argmin}_{\boldsymbol{\theta}\in\Lambda^M} V_{\mathrm{pen}}(\boldsymbol{\theta})$ où*

$$V_{\mathrm{pen}}(\boldsymbol{\theta}) := H_{\mathrm{pen}}(\boldsymbol{\theta}) + 30\sigma^2\mathcal{K}\theta. \tag{8.26}$$

Alors pour tout $x > 0$, avec probability plus grande que $1 - \exp(-x)$, nous avons

$$|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}} - \boldsymbol{f}|_2^2 \leq \min_{j=1,...,M}\left(|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 60\sigma^2 \log\frac{1}{\pi_j}\right) + 30\sigma^2 x. \tag{8.27}$$

De plus,

$$\mathbb{E}|\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\theta}}_{\boldsymbol{\pi}}} - \boldsymbol{f}|_2^2 \leq \mathbb{E}\min_{j=1,...,M}\left(|\hat{\boldsymbol{\mu}}_j - \boldsymbol{f}|_2^2 + 60\sigma^2 \log\frac{1}{\pi_j}\right). \tag{8.28}$$

La mesure de probabilité a priori $\boldsymbol{\pi} = (\pi_j)_{j=1,...,M}$ est déterministe et ne peut pas dépendre des données $\mathbf{y} = (Y_1, ..., Y_n)^T$. La seule différence entre la fonction (8.5) et la fonction minimisé dans (8.26) est le terme

$$\sigma^2\mathcal{K}\theta.$$

Ce terme nous permet de donner des poids différents aux estimateurs candidats $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$ avec la mesure de probabilité $(\pi_j)_{j=1,...,M}$ basé sur une connaissance préalable sur les caractéristiques des estimateurs $\hat{\boldsymbol{\mu}}_1, ..., \hat{\boldsymbol{\mu}}_M$. Par exemple, si les estimateurs sont des projecteurs orthogonaux on peut définir une mesure de probabilité a priori qui décroit avec le rang des projecteurs [92, 93]. Le même terme est utilisé [69] tandis que [33] utilise la divergence de Kullback-Leibler de $\boldsymbol{\theta}$ par rapport $\boldsymbol{\pi}$.

8.2 Résumé des différents chapitres

Les chapitres peuvent être lues de manière indépendante. Voici un résumé succin de chaque chapitre et de leur contribution principale.

- Le chapitre 2 étudie le problème d'agrégation de fonctions déterministes pour l'estimation de densité en perte L^2. Les résultats principaux de ce chapitre sont les inégalités d'oracle données dans les Théorèmes 2.6 et 2.8.

- Le chapitre 3 étudie le problème d'agrégation d'estimateurs affines en régression à design fixe. Les estimateurs sont dépendants des données utilisées pour l'agrégation. Le résultat du chapitre 3 est l'inégalité d'oracle donné dans le Théorème 3.1.

- Dans le chapitre 4, nous construisons un estimateur qui agrège les estimateurs Lasso sur le chemin de régularisation du Lasso. Cet estimateur est presque aussi performant que le meilleur estimateur Lasso, cf. Théorème 4.3.

- Le chapitre 5 lie les deux domaines des statistiques étudiés dans cette thèse: l'agrégation d'estimateurs et la régression sous contrainte de forme.

- Le chapitre 6 étudie l'estimateur des moindres carrés en régression sous contrainte de forme. Le résultat principal de ce chapitre est que les inégalités d'oracle du chapitre 5 sont également satisfaites par l'estimateur des moindres carrées.

- Enfin, dans le chapitre 7 nous construisons des ensembles de confiance dans le contexte de la régression sous contrainte de forme. Le chapitre 7 prouve l'existence d'ensembles de confiance qui capturent la vrai fonction avec grande probabilité et dont le diamètre est de l'ordre de la vitesse minimax, cf. Theorems 7.2, 7.3, 7.9 and 7.11.

8.3 Mise en perspective et notes bibliographiques

Les premiers résultats sur l'agrégation dans un contexte statistique sont apparus dans Nemirovski [83], Catoni [25], Yang [106] et Tsybakov [98]. Ces travaux pionniers étudient trois différents problème d'agrégation.

- Pour l'agrégation de type sélection de modèles, le but est d'imiter la performance de la meilleure fonction dans le dictionnaire. Les résultats pour ce problème ont été obtenus dans [106, 25, 98, 64, 70, 58, 4, 35, 90, 32, 33], Chapters 2 and 3.

- Pour le problème d'agrégation convexe, le but est d'imiter la performance de la meilleure combinaison convexe des fonctions du dictionnaire [98, 89, 90, 96]. La Proposition 3.10 du chapitre 3 donne un résultat d'agrégation convexe pour les estimateurs affines.

- Pour le problème d'agrégation linéaire, le but est d'imiter la meilleure fonction dans l'espace vectoriel engendré par les fonctions du dictionnaire. [98, 89, 90, 96].

Des travaux plus récents étudient le problème d'agrégation parcimonieuse et celui de l'agrégation parcimonieuse-convexe, cf. [71, 92, 93, 96].

Cette thèse se concentre sur le problème d'agrégation de type sélection de modèle. L'estimateur pénalisé étudié dans les chapitres 2 et 3 est similaire à la procédure de Q-agrégation proposé par Rigollet [90] et Dai et al. [32]. L'agrégation d'estimateurs affines utilisant la Q-agrégation a été étudié précédemment dans Dai et al. [33].

Leung and Barron [70] ont donné le premier résultat d'agrégation d'estimateurs linéaires, où il est nécessaire de prendre en compte la dépendance entre les estimateurs dans le dictionnaire et les données utilisées pour la phase d'agrégation. Ces résultats ont été plus tard généralisés dans Dalalyan and Salmon [35], Dai et al. [33] et dans le chapitre 3 de la présente thèse. A notre connaissance, le Théorème 4.2 du chapitre 4 est le premier résultat d'agrégation d'estimateurs non linéaires où les estimateurs non linéaires sont construits à partir des mêmes données que celles utilisées pour la phase d'agrégation.

Le chapitre 5 explique comment les méthodes d'agrégation peuvent être utilisées pour produire des inégalités d'oracle exactes dans le cadre de la régression à contrainte de forme, généralisant les résultats de Guntuboyina and Sen [50], Chatterjee et al. [27] et Chatterjee et al. [28]. Ces papiers ont d'abord étudié la vitesse quasi-paramétrique qui apparaît si la fonction de régression inconnue possède des propriétés de basse dimension, cf. les chapitres 5 et 6 pour des résultats rigoureux et une discussion plus approfondie à propos de ces propriétés de basse dimension. A notre connaissance, le chapitre 7 donne les premiers résultats sur la construction d'ensembles de confiance adaptatifs dans ce cadre.

8.4 Remerciements

Je tiens d'abord à remercier Sacha, mon directeur de thèse qui a accepté de m'encadrer au début ce ces trois années intenses. Merci pour sa confiance, les problèmes qu'il a su choisir (la réussite d'une thèse repose beaucoup sur le choix des problèmes !), ses idées fécondes, sa patience et sa clarté, les interactions que nous avons eu pendant ces trois années m'ont appris énormément. J'espère que nous aurons encore longue collaboration, en plus des quelques projets planifiés pour les mois qui viennent.

J'aimerais exprimé une grande reconnaissance aux rapporteurs, Richard et Bodhi, qui ont accepté d'écrire un rapport sur ce manuscrit. Merci également à Richard pour nos discussions sur les ensembles de confiance, et à Bodhi pour son invitation à Columbia et nos échanges sur la régression sous contrainte de forme. Mes remerciements vont ensuite au jury de thèse. Arnak, pour les nombreuses discussions que nous avons eu dans la salle à café de l'ENSAE qui m'ont tant appris, en particulier sur les statistiques en hautes dimensions. Richard, pour ces discussions fructueuses sur les ensembles de confiance qui a abouti sur le chapitre 7 de cette thèse. Philippe, pour tes nombreux travaux qui ont inspiré la plupart des résultats de cette thèse, pour tes précieux conseils sur la suite de mon parcours, et j'espère que nos interactions s'amplifieront sur la côté Est dans les prochaines années! Vladimir, merci encore d'avoir accepté de participer à ce jury de thèse, tes notes de St-Flour ont été précieuses de nombreuses fois pendant ces trois ans.

Merci aussi à tous l'équipe de laboratoire de stats de l'ENSAE : Pierre, Victor-Emmanuel, Alexander, Léna, Nicolas, Vincent, Arnak, Edwin, Gérard, Olga, Guillaume, Hilmar, The Tien, Vianney, Judith, Medhi et Anna pour les nombreux

moments conviviaux passés à l'ENSAE.

Enfin, merci à mes parents pour leur soutien infaillible tout au long de ces trois années. Merci à L., M., A., Q., S., H., R., D., C., N., M., S. pour leur bonne humeur et tous les moments passés ensemble – puissent ces relations pendant de nombreuses années !

Résumé

Deux sujet sont traités dans cette thèse: l'agrégation d'estimateurs et la régression sous contrainte de formes.

La régression sous contrainte de forme étudie le problème de régression (trouver la fonction qui représente un nuage de points), avec la contrainte que la fonction en question possède une forme spécifique. Par exemple, cette fonction peut être croissante ou convexe: ces deux contraintes de forme sont les plus étudiés. Nous étudions en particulier deux estimateurs: un estimateur basé sur des méthodes d'agrégation et un estimateur des moindres carrés avec une contrainte de forme convexe. Des inégalités d'oracle sont obtenues, et nous construisons aussi des intervalles de confiance honnêtes et adaptatifs.

L'agrégation d'estimateurs est le problème suivant. Lorsque plusieurs méthodes sont proposes pour le même problème statistique, comment construire une nouvelle méthode qui soit aussi performante que la meilleure parmi les méthodes proposées? Nous étudierons ce problème dans trois contextes: l'agrégation d'estimateurs de densité, l'agrégation d'estimateurs affines et l'agrégation sur le chemin de régularisation du Lasso.

Summary

This PhD thesis studies two fields of Statistics: Aggregation of estimators and shape constrained regression.

Shape constrained regression studies the regression problem (find a function that approximates well a set of points) with an underlying shape constraint, that is, the function must have a specific "shape". For instance, this function could be nondecreasing of convex: These two shape examples are the most studied. We study two estimators: an estimator based on aggregation methods and the Least Squares estimator with a convex shape constraint. Oracle inequalities are obtained for both estimators, and we construct confidence sets that are adaptive and honest.

Aggregation of estimators studies the following problem. If several methods are proposed for the same task, how to construct a new method that mimics the best method among the proposed methods? We will study these problems in three settings: aggregation of density estimators, aggregation of affine estimators and aggregation on the regularization path of the Lasso.